THE UNIVERSITY OF CHICAGO

EXPONENTIAL SERIES APPROACHES FOR NONPARAMETRIC GRAPHICAL
MODELS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY

ERIC JANOFSKY

CHICAGO, ILLINOIS

JUNE 2015

To grammy

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost I thank my advisor John Lafferty. His depth and breadth of knowledge has had a profound influence on me since I first took his statistical machine learning elective in 2012. His assistance, patience and encouragement over the last several years has been invaluable to the completion of this thesis. I also thank my committee members Matthew Stephens and Lek-Heng Lim for their constructive input, as well as the rest of the faculty of the Department of Statistics for the outstanding education I received in the last five years, and before that during my undergraduate studies.

I thank my family for their love and support, especially my wife Liz. She has acted as a source of strength and inspiration throughout this journey, all while pursuing her own doctorate. I couldn't have done it without her.

ABSTRACT

Markov Random Fields (MRFs) or undirected graphical models are parsimonious representations of joint probability distributions. Variables correspond to nodes of a graph, with edges between nodes corresponding to conditional dependencies. For a *pairwise MRF*, the joint density factorizes as a product over edges of the graph. This thesis studies high-dimensional, continuous-valued pairwise MRFs. We are particularly interested in approximating pairwise densities whose logarithm belongs to a Sobolev space. For this problem we propose the method of exponential series [Crain, 1974; Barron and Sheu, 1991], which approximates the log density by a finite-dimensional exponential family with the number of sufficient statistics increasing with the sample size.

We consider two approaches to estimating these models. The first is regularized maximum likelihood. This involves optimizing the sum of the log-likelihood of the data and a sparsity-inducing regularizer. We provide consistency and edge selection guarantees for this method. We then propose a variational approximation to the likelihood based on tree-reweighted, nonparametric message passing. This approximation allows for upper bounds on risk estimates, leverages parallelization and is scalable to densities on hundreds of nodes. We show how the regularized variational MLE may be estimated using a proximal gradient algorithm. We demonstrate our method's efficacy in density estimation and model selection in comparison to other approaches in the literature using simulated data and MEG signal data.

We then consider estimation using regularized score matching. This approach uses an alternative scoring rule to the log-likelihood, which obviates the need to compute the normalizing constant of the distribution. For general continuous-valued exponential families, we provide parameter and edge consistency results. As a special case we detail a new approach to sparse precision matrix estimation which has statistical performance competitive with the graphical lasso [Yuan and Lin, 2007] and computational performance competitive with the state-of-the-art glasso algorithm [Friedman et al., 2008]. We then describe results for model

selection in the nonparametric pairwise model using exponential series. The regularized score matching problem is shown to be a convex program; we provide scalable algorithms based on consensus alternating direction method of multipliers (ADMM, [Boyd et al., 2011]) and coordinate-wise descent. We use simulations to compare our method to others in the literature as well as the aforementioned TRW estimator.

CHAPTER 1

INTRODUCTION

Density estimation is one of the fundamental tools in statistics and machine learning [Silverman, 1986]. Nonparametric density estimators are used for prediction, goodness-of-fit testing [Fan, 1994; Bickel and Rosenblatt, 1973], generative models for classification [Fix and Hodges, 1989; John and Langley, 1995], inferring independence and conditional independence, estimating statistical functionals [Beirlant et al., 1997; Póczos et al., 2012], as well as exploratory data analysis and visualization. In many fields including the biological and social sciences, information technology and machine learning, it is now commonplace to analyze large data sets with complex dependencies between many variables. This includes gene expression levels measured using microarrays, brain activity measurements from fMRI or MEG technology, prices of financial instruments, and the activity of individuals on social media web sites. The use of nonparametric density estimators is limited, however, by the curse of dimensionality: with even moderate sample size, high-dimensional space will invariably have large regions where the data is sparse, leading to uninformative predictions. As such nonparametric estimators typically have poor risk guarantees in high-dimensions. Furthermore, many methods suffer from a computational curse of dimensionality and heuristics or approximations to tune the models become necessary.

In this work we consider nonparametric estimation of *pairwise densities*, a class of densities intimitely tied to undirected graphical models. The undirected graphical model or Markov Random Field [Jordan, 2004; Lauritzen, 1996] is a well-studied framework for representing joint dependence structures of random variables. An undirected graph $G = (V, E)$ consists of a vertex set $V = \{1, \ldots, d\}$ corresponding to the elements of the random vector $X = (X_1, \ldots, X_d)$, and an edge set $E \subseteq V \times V$. Each edge $e \in E$ is an unordered pair of elements $j, k \in V$, $e = (j, k)$. For any subset $A \subseteq V$, we define the subset $\{X_A : X_i, i \in A\}$. Furthermore, for sets A, B, C we write $X_A \perp\!\!\!\perp X_B | X_C$ to mean X_A and X_B are independent conditional on X_C. The random vector X is *Markov* with respect to the graph $G = (V, E)$

if for every $j, k \in V$, $X_j \perp\!\!\!\perp X_k | (X_l : l \neq j, k)$ if and only if $(i, j) \notin E$.

The fundamental theorem of undirected graphical models is the Hammersley-Clifford theorem [Dobruschin, 1968], which states that if the density of X $p(x)$ is positive, then the following are equivalent:

1. X is Markov with respect to the graph G;

2. The density of X, $p(x)$ can be factorized over the cliques of G:

$$p(x) = \prod_{C \in cl(G)} \psi_C(x_C), \tag{1.1}$$

where $x_C = \{x_i : i \in C\}$.

This thesis considers nonparametric estimation of the pairwise graphical model for continuous-valued data, where the joint density can be further factored into a product of potential functions over edges:

$$p(x) = \prod_{i \in V} \psi_i(x_i) \prod_{(j,k) \in E} \psi_{jk}(x_j, x_k). \tag{1.2}$$

Pairwise graphical models have been used extensively in modeling discrete data. The Ising model [Ising, 1925] for $\{0, 1\}$-valued variables has the density

$$p(x) = \exp \left\{ \sum_{i \in V} \theta_i x_i + \sum_{(j,k) \in E} \theta_{jk} x_j x_k - Z(\theta) \right\},$$

which can be seen as a pairwise graphical model with $\psi_i = \exp\{\theta_i x_i\}$ and $\psi_{jk} = \exp\{\theta_{jk} x_j x_k\}$ and $Z(\theta)$ a normalizing constant. The Ising model can be generalized to discrete variables with more than two levels, but there are a finite possibilities for pairwise discrete potentials with finite number of levels.

The class of continuous-valued pairwise models is considerably more complex than discrete ones, as the potential functions $\{\psi_i, \psi_{jk}\}$ could be any positive-valued functions such

that p integrates to 1. The class of continuous pairwise graphical models includes some familiar models, which we describe below.

Example 1.0.1. Gaussian graphical model

Let $X \in \mathbb{R}^d$ be a Gaussian-distributed random variable with mean μ and covariance matrix $\mathbb{E}\big((X - \mu)(X - \mu)^\top\big) = \Sigma \succ 0$. Denote $\Omega = \Sigma^{-1}$. Then X has density

$$
\begin{aligned}
p(x) \quad &= \quad \sqrt{\frac{|\,\Omega\,|}{(2\pi)^d}} \exp\left\{ -\frac{1}{2}(x - \mu)^\top \Omega (x - \mu) \right\} \\
&\propto \quad \prod_{i \in V} \exp\left\{ -\frac{1}{2}\Omega_{ii}x_i^2 + (\Omega\mu)_i x_i \right\} \\
&\qquad \times \prod_{(i,j) \in E} \exp\left\{ -\frac{1}{2}\Omega_{ij}x_i x_j \right\}.
\end{aligned}
$$

From the factorization above, it can be seen that the Gaussian graphical model is not only a Markov random field, but also belongs to the pairwise class of densities. Following from (1.1), observing that the Gaussian density is positive over $\mathbb{R}^d$, two Gaussian variables X_i, X_j are conditionally independent given the others if and only if $(\Sigma^{-1})_{ij} = 0$.

Example 1.0.2. Gaussian copula graphical model

Let X be Gaussian distributed with mean μ and covariance matrix Σ, and suppose that for each $i \in V$, $Y_i = g_i(X_i)$ where g_i is some monotonic increasing, smooth function. Suppose further that the g_i are centered and scaled so that $\mathbb{E}[g_i(X_i)] = \mathbb{E}[X_i]$ and $\mathrm{var}[g_i(X_i)] = \mathrm{var}[X_i]$. Write $f = g^{-1}$. Denote the vector of functions $f = (f_1, f_2, \ldots, f_d)$ and their derivatives by $f' = (f_1', f_2', \ldots f_d')$. After applying a change of variables to the Gaussian

density, we find that Y has density

$$
\begin{aligned}
p(y) &= \sqrt{\frac{|\,\Omega\,|}{(2\pi)^d}} \exp\left\{ -\frac{1}{2}(f(y)-\mu)^\top \Omega (f(y)-\mu) \right\} \prod_{i\in V} f_i'(y_i) \\
&\propto \prod_{i\in V} f_i'(y_i) \exp\left\{ -\frac{1}{2}\Omega_{ii} f_i(y_i)^2 + (\Omega\mu)_i f_i(y_i) \right\} \\
&\quad \times \prod_{(i,j)\in E} \exp\left\{ -\frac{1}{2}\Omega_{ij} f_i(y_i) f_j(y_j) \right\}.
\end{aligned}
$$

Thus the Gaussian copula density is also a Markov random field and has a pairwise factorization. Following from (1.1), two variables Y_i, Y_j are conditionally independent given the others if and only if $\Omega_{ij} = 0$. Gaussian copulas have been used extensively in finance and risk management [Cherubini et al., 2004] for their ability to model dependence between many variables which are (marginally) non-Gaussian.

Example 1.0.3. Forests

A *tree T* is an undirected graph where each pair of vertices is connected by exactly one simple path. Equivalently, a tree is a connected graph with no cycles. A *forest* is an undirected graph where each pair of vertices is connected by at no more than one simple path. Equivalently, a forest is a graph with no cycles. A *spanning tree T* of a connected graph G is a tree containing the vertices of G and a subset of the edges of G. A *spanning forest* of a graph G is a graph consisting of a spanning tree for each connected component of G. In the sequel will use the term spanning tree to refer to a spanning tree or forest unambiguously whether or not G is connected.

A tree or forest must have cliques of size no more than two. Thus, from the Hammersley-Clifford theorem, a tree density has the factorization

$$
p(x) = \prod_{i\in V} \psi_i(x_i) \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j). \tag{1.3}
$$

It follows that all tree distributions have pairwise densities. In particular, a tree can always

be factorized in the form

$$p(x) = \prod_{i \in V} p_i(x_i) \prod_{(i,j) \in E} \frac{p_{ij}(x_i, x_j)}{p_i(x_i) p_j(x_j)}, \qquad (1.4)$$

where $\{p_i\}$ are the set of univariate densities and $\{p_{ij}\}$ are the set of bivariate densities of the joint distribution p.

1.1 Previous Work

Little work has been done in studying the nonparametric estimation of pairwise densities. [Gu, 2002, 1993] considered log-ANOVA density estimation, where the log-density can be factored into low-order terms, pairwise densities being a special case. They suggest an estimator for the log-density:

$$\min_{\eta} \left\{ -\frac{1}{n} \sum_{k=1}^{n} \eta(X^k) + \log \int e^{\eta} + \lambda \|\eta\|_{\mathcal{H}}^2 \right\}, \qquad (1.5)$$

such that η has a given pairwise factorization, where $\|\cdot\|_{\mathcal{H}}$ is a norm in a reproducing kernel Hilbert space (RKHS); the resulting density estimate is proportional to $e^{\eta(x)}$. Due to the representer theorem [Kimeldorf and Wahba, 1971], this becomes a finite-dimensional optimization problem. However, due to the difficulty of computing $\log \int e^{\eta}$ it can only be used in low-dimensional problems, such as dimension up to 3. This work also assumes the ANOVA factorization structure is known. [Jeon and Lin, 2006] proposed a smoothing spline estimator based on minimizing the *Bregman score* (4.8) for log-ANOVA densities, and applied it to undirected graphical model estimation. This solves the problem

$$\min_{\eta} \left\{ \frac{1}{n} \sum_{k=1}^{n} e^{-\eta(X^k)} + \int \eta(y)\rho(y)dy + \lambda \|\eta\|_{\mathcal{H}}^2 \right\}, \qquad (1.6)$$

5

for some given density ρ; the resulting density estimate is proportional to $\rho(x)e^{\eta(x)}$. The authors show that solving this problem only requires calculation of one-dimensional integrals, and is thus more scalable. This approach has some limitations. The issue of computing the normalizing constant remains; this is a problem for inference and for choosing the smoothing parameter λ when using cross-validation to minimize the KL risk. Also, their procedure for graph selection is a heuristic. The theoretical properties of this estimator are not yet known.

When the graph is assumed to be a forest (Example 1.0.3), density estimation and structure learning was considered in [Liu et al., 2011]. Due to the tree entropy factorization (2.19), learning a forest density can be done in two steps: estimating the univariate and bivariate marginals, and learning the graph structure. For the first task, they estimate marginals using kernel density estimation. For the second task, they use a nonparametric estimator of mutual information (2.11), and then estimate the maximum likelihood forest using Kruskal's algorithm (Figure 3.4), with weights corresponding to mutual information between edges. They show consistency guarantees for their approach in terms of KL risk and graph selection.

There has been a large amount of work on parametric pairwise models. For learning Gaussian models with sparse precision matrix, the graphical lasso, or the L_1-regularized maximum likelihood is the most popular approach [Yuan and Lin, 2007]; this solves the problem

$$\min_{\Omega \succ 0} \left\{ \text{trace}(\Omega\widehat{\Sigma}) - \log|\Omega| + \lambda\|\Omega\|_1 \right\}, \tag{1.7}$$

$\widehat{\Sigma}$ being the sample covariance matrix $\widehat{\Sigma} = \frac{1}{n}\sum_{k=1}^n (X^k)(X^k)^\top$ and $\|\Omega\|_1 = \sum_{i,j}\left|\Omega_{ij}\right|$. The glasso algorithm solves the resulting problem using block-coordinate descent [Banerjee et al., 2008; Friedman et al., 2008]. The graphical lasso is known to have good properties in terms of parameter and structure selection consistency [Rothman et al., 2008; Ravikumar et al., 2011]. There exist other approaches for sparse estimation of Ω such as the the graphical Danzig selector [Yuan, 2010], and CLIME [Cai et al., 2011]. The parallel lasso [Meinshausen

and Bühlmann, 2006] infers graph structure in a Gaussian graphical model without direct estimation of the covariance or precision matrix by running a sequence of neighborhood lasso fits in parallel. For a node i, they estimate the neighborhood of i by solving

$$\widehat{\beta}^i = \operatorname*{argmin}_{\beta:\beta_i=0} \left\{ \frac{1}{2n} \sum_{k=1}^{n} \left(X_i^k - \beta^\top X^k \right)^2 + \lambda \|\beta\|_1 \right\}. \tag{1.8}$$

The corresponding neighborhood of i is the support of $\widehat{\beta}^i$. They show consistency of neighborhood estimates, which may then be aggregated to form an edge set. The precision matrix can then be fit by estimating the Gaussian likelihood subject to sparsity constraints on the precision matrix. [Liu et al., 2012a] proposes the SKEPTIC estimator for structure learning of the semiparametric Gaussian copula model. The estimator plugs in a matrix of rank correlations (Kendall's τ or Spearman's ρ) into the graphical lasso (1.7); they show this estimator achieves the parametric rates for edge selection and parameter estimation.

This thesis includes several contributions to the literature. In Chapter 2 we introduce the exponential series approximation to pairwise densities. We propose an estimator for pairwise densities based on regularized maximum likelihood estimation of a particular exponential family whose sufficient statistics are basis elements. We use a method for edge selection using convex regularization, and provide risk and model selection guarantees in Section 2.5. While the exact problem is in general not tractable, in Chapter 3 we propose a convex variational upper bound on the likelihood based on a nonparametric tree-reweighted relaxation [Wainwright et al., 2005, 2003], which can be computed efficiently and in parallel. Our method provides an upper bound on the normalizing constant, guaranteeing an upper bound on risk estimates. The approximation leads to a natural variational maximum likelihood estimator, as well as an approach for marginalization. We train our method using a projected gradient algorithm, which can effortlessly be scaled to relatively sparse graphs on hundreds of nodes. In Section 3.6 we compare our method to several other approaches to large-scale density estimation, including the graphical lasso, mixtures of Gaussians with

the EM algorithm, and kernel forest density estimation. We demonstrate our method by estimating the graph from an MEG neuroimaging dataset.

In Chapter 4 we consider a different approach to estimation and graph selection using an alternative scoring rule to the log-likelihood. It is based on minimizing the log-gradient between the model distribution and data distribution, or equivalently minimizing the *Fisher divergence*. This method obviates the need for computing a normalizing constant. We show that the optimization amounts to a second-order cone program, and provide two types of scalable algorithms specially tailored to the problem. Our method, which we denote QUASR for **Qua**dratic **S**coring and **R**egularization, produces parameter and graph selection consistency for general pairwise exponential families with only weak regularity conditions. En route, we derive a new method for sparse precision matrix estimation which performs competitively with the regularized MLE (1.7). Finally, we show how this approach produces graph selection guarantees for the pairwise nonparametric model when the sufficient statistics are basis elements.

1.2   Notation and Preliminaries

Throughout this thesis we assume we are given independent and identically distributed data $X^1, X^2, \ldots, X^n$, where $X^k = (X_1^k, \ldots, X_d^k)$, drawn from the density $p(x)$ with respect to a reference measure $\nu(x)$. In the context of density estimation, we assume the unknown log density $f = \log p$ belongs to the Sobolev space of functions on $[0,1]^d$, W_r^2, so that for any multi-index α with $\mid \alpha \mid \leq r$,

$$f^{(\alpha)} := \frac{\partial^{(\alpha)} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \tag{1.9}$$

has bounded $L_2(\nu)$ norm:

$$\|f^{(\alpha)}\| := \int_{\mathcal{X}} \mid f^{(\alpha)} \mid^2 \nu(dx) < \infty. \tag{1.10}$$

8

This implies that p is bounded away from zero and infinity and that $p(x)$ has bounded support.

We use the asymptotic notations $O(\cdot)$, $o(\cdot)$, $\Omega(\cdot)$ and $\asymp$. For two functions $f(n), g(n)$, $f = O(g)$ if $f/g \le c$ for a constant $c > 0$ as $n \to \infty$; $f = o(g)$ if $f/g \to 0$ as $n \to \infty$; $f = \Omega(g)$ if $f > cg$ for some $c > 0$ as $n \to \infty$; $f \asymp g$ if $f = cg$ for some $c > 0$ as $n \to \infty$. We say that $f = O_p(g)$ if $f = O(g)$ with probability approaching one as $n \to \infty$.

Let $\{\phi_k, \phi_{k'l}, k, k', l = 1, 2, \ldots\}$ be a tensor product basis for $L_2[0,1]^2$, so that $\phi_{kl} = \phi_k \phi_l$. We suppose this basis is uniformly bounded and orthonormal. Consider the density p having a pairwise factorization 1.2, so that $f = \log p$ can be expressed with the basis expansion

$$f(x) = f_0(x) + \theta_0^* + \sum_{i,j \in V} \sum_{k,l=1}^{\infty} (\theta^*)_{ij}^{kl} \phi_k(x_i)\phi_l(x_j) + \sum_{i \in V} \sum_{k=1}^{\infty} (\theta^*)_i^k \phi_k(x_i), \qquad (1.11)$$

$\exp f_0$ is some base measure which has the same pairwise factorization as p. We will take $f_0 = 0$, but our results also apply whenever f_0 has the same smoothness assumptions as f. θ, ϕ represent the parameters and sufficient statistics vectorized. θ_e, θ_v denote the vectors of edge and vertex parameters, respectively. For a $f \in W_r^2$, for any $i, j \in V$ and $r_i + r_j = r$, we additionally have that

$$\sum_k (\theta_i^{*k})^2 k^{2r} = C_1 < \infty, \qquad (1.12)$$

$$\sum_{k,l} (\theta_{ij}^{*kl})^2 k^{2r_i} l^{2r_j} = C_2 < \infty. \qquad (1.13)$$

This implies that $\left|(\theta^*)_i^k\right| = o(k^{-r-1/2})$ and $\left|(\theta^*)_{ij}^{kl}\right| = o(k^{-r_i-1/2} l^{-r_j-1/2})$.

We denote Π to be the set of densities on $[0,1]^d$. For risk analysis in Chapter 2 we use the relative entropy, also known as the Kullback-Leibler divergence:

$$\mathrm{KL}(p \mid \widehat{p}) = \int_{\mathcal{X}} p(x) \log \left(\frac{p(x)}{\widehat{p}(x)}\right) d\nu(x) = \mathbb{E}_p \left[\log \left(\frac{p(X)}{\widehat{p}(X)}\right)\right]. \qquad (1.14)$$

9

It can be shown that $\text{KL}(p \mid \widehat{p}) \geq 0$ and equals zero only if $p = \widehat{p}$ ν-almost everywhere. Relative entropy is a natural risk measure for density estimation. It is invariant to invertible changes of variables, and shares a natural connection to maximum likelihood. Convergence in KL is strong in the sense that it implies convergence in several other risk measures. In particular, define the L_1, Hellinger and Total Variation distances as follows:

$$D_1(p \mid \widehat{p}) = \int_{\mathcal{X}} \mid p(x) - \widehat{p}(x) \mid d\nu(x), \tag{1.15}$$

$$D_H(p \mid \widehat{p}) = \int_{\mathcal{X}} \left| p(x)^{1/2} - \widehat{p}(x)^{1/2} \right|^2 d\nu(x), \tag{1.16}$$

$$D_{TV}(p \mid \widehat{p}) = \sup_{A \in \mathcal{X}} \left| \int_A p(x) d\nu(x) - \int_A \widehat{p}(x) d\nu(x) \right|, \tag{1.17}$$

By Pinsker's inequality and [Reiss, 1989; Kullback, 1967], we have

$$\text{KL}(p \mid \widehat{p}) \geq \frac{D_1(p \mid \widehat{p})^2}{2}, \tag{1.18}$$

$$\text{KL}(p \mid \widehat{p}) \geq D_H(p \mid \widehat{p})^2, \tag{1.19}$$

$$\text{KL}(p \mid \widehat{p}) \geq 2 D_{TV}(p \mid \widehat{p})^2. \tag{1.20}$$

CHAPTER 2

EXPONENTIAL FAMILIES AND THE EXPONENTIAL
SERIES REGULARIZED MLE

2.1 Exponential Series Approximation

Consider an approximation to $f = \log p$ (1.11) by truncating the basis expansion in the following way:

$$\log p_\theta(x) = \sum_{i=1}^{d} \sum_{k=1}^{m_1} \theta_i^k \phi_k(x_i) + \sum_{i,j \in V} \sum_{k=1}^{m_2} \sum_{l=1}^{m_2} \theta_{ij}^{kl} \phi_{kl}(x_i, x_j) - Z(\theta), \tag{2.1}$$

This approximation is an exponential family with sufficient statistics corresponding to the basis functions $\{\phi_k(x_i), i \in V, k \leq m_1\}$ and $\{\phi_k(x_i)\phi_l(x_j), i, j \in V, k, l \leq m_2\}$. $Z(\theta)$ is chosen so the density integrates to one. The idea of representing a density as an exponential expansion was first used for goodness of fit testing in [Neyman, 1937]. Nonparametric estimation of univariate distributions using exponential series has been studied previously in [Crain, 1974, 1977; Barron and Sheu, 1991]. For simplicity we assume that each univariate component is truncated after m_1 terms and bivariate components after m_2 terms. In practice we could attempt to vary truncation for each variable, though this could be unwieldy for large problems. When not ambiguous we will write $\log p_\theta(x) = \langle \theta, \phi(x) \rangle - Z(\theta)$. Given n i.i.d. samples $X^1, \ldots, X^n$, the *exponential series MLE estimator* finds the regularized maximum likelihood estimator of θ:

$$\widehat{\theta} \ := \ \underset{\theta}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{r=1}^{n} \log p_\theta(X^r) + \lambda \mathcal{R}(\theta) \right\}, \tag{2.2}$$

$$= \ \underset{\theta}{\operatorname{argmin}} \left\{ -\mathcal{L}(\theta) + \lambda \mathcal{R}(\theta) \right\}. \tag{2.3}$$

$\mathcal{R}$ is a convex regularizer; we discuss our particular choice of regularization in Section 2.5.1. The resulting density estimate is $p_{\widehat{\theta}}(x)$. The fact that θ depends on the truncation parameters

m_1, m_2 and regularization parameter λ is left implicit.

Exponential series is a natural approach for the estimation of pairwise densities. The product factorization of pairwise densities can be expressed naturally by exponential series. Indeed, many common parametric graphical models are exponential families. Each truncated series forms an exponential family. We detail exponential families and the regularized maximum likelihood in the sequel.

2.2 Exponential Families

This section presents background on exponential families and their use in graphical modeling. For an in-depth treatment, see [Brown, 1986; Wainwright and Jordan, 2008].

Consider the random vector $X = (X_1, \ldots, X_d)$ taking values on the support $\mathcal{X} = \bigotimes_{i=1}^{d} \mathcal{X}_i$. The *exponential family* with sufficient statistics $\phi(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_M(x))$ is the family of probability distributions

$$\left\{ p_\theta : p_\theta(x) = \exp\left\{ \langle \theta, \phi(x) \rangle - Z(\theta) \right\}, \theta \in \mathcal{P} \right\}, \tag{2.4}$$

where $\langle \theta, \phi(x) \rangle = \sum_{a=1}^{M} \theta_a \phi_a(x)$. $Z(\theta)$ is called the *log-partition function*, and is given by

$$Z(\theta) = \log \int_{\mathcal{X}} \exp\left\{ \langle \theta, \phi \rangle \right\} \nu(dx), \tag{2.5}$$

and ensures the density $p_\theta(x)$ integrates to one. The family is indexed by the parameters θ, called the *natural parameters* belonging to the space $\mathcal{P} = \{\theta : Z(\theta) < \infty\}$.

An exponential family is *minimal* if there is no choice of parameters $\theta \neq 0$ such that $\langle \theta, \phi(X) \rangle = C$ ν-a.e., where C is a constant. If a family is minimal, there is a bijection between the natural parameter space $\mathcal{P}$ and the densities belonging to the family. An exponential family is *regular* if $\mathcal{P}$ is an open set.

Example 2.2.1 (Gaussian exponential family). Consider the Gaussian density with mean μ

and covariance $\Sigma \succ 0$. The Gaussian family is an exponential family with sufficient statistics $\{x, xx^\top\}$ and natural parameters $(\theta_1, \theta_2) = \{\Omega\mu, -\frac{1}{2}\Omega\}$. Since $\Sigma \succ 0$, the natural parameter space is

$$\mathcal{P} = \{(\theta_1, \theta_2) \in \mathbb{R}^d \times \mathbb{R}^{d \times d} : \theta_2 \prec 0\}.$$

The space of negative definite matrices is an open convex set, so it follows that $\mathcal{P}$ is convex and the Gaussian family is regular. Furthermore, linear independence of monomials implies that x and $xx^\top$ are linearly independent, and so the Gaussian family is also minimal.

Remark 2.2.1 (Exponential series). Consider the exponential series family of (2.1). When $\{\phi_k, \phi_{kl}\}$ is an orthogonal basis which satisfies the Haar condition, the exponential series family corresponds to a minimal exponential family. Furthermore, since the exponential series are defined have compact support $[0, 1]^d$ and the sufficient statistics ϕ are bounded above and below, any choice of $\theta \in \mathbb{R}^M$ will produce a valid $Z(\theta) < \infty$; in other words $\mathcal{P} = \mathbb{R}^M$, which is a convex and open set, so the exponential series family is regular.

2.3 Mean Parametrization

An exponential family is parametrized by its so-called *natural parameters* θ (2.4). Alternatively, an exponential family may also be characterized by a vector of *mean parameters* μ. The connection between these seemingly disparate entities will be described in Lemma 2.3.1.

Let p be any density on $\mathcal{X}$ with respect to ν. We define the mean parameter μ_a corresponding to the sufficient statistic ϕ_a by

$$\mu_a = \int p(x)\phi_a(x)d\nu(x) = \mathbb{E}_p[\phi_a]. \tag{2.6}$$

Consider the set of vectors that correspond to the moments of some distribution: $\mathcal{M} = \{\mu : \exists p \in \Pi, \mu_a = \mathbb{E}_p[\phi_a], \forall a = 1, \ldots, M\}$. In particular, the elements of $\mathcal{M}$ need not correspond to mean parameters of an exponential family. Furthermore, $\mathcal{M}$ is a convex set. To see this,

let $\mu_1, \mu_2 \in \mathcal{M}$ be mean parameters corresponding to two distributions p_1, p_2 with respect to ν. Then

$$\lambda\mu_1 + (1 - \lambda)\mu_2 = \mathbb{E}_{\lambda p_1 + (1-\lambda)p_2}[\phi].$$

For discrete variables, one can further show that $\mathcal{M}$ is a convex polytope [Wainwright and Jordan, 2008]. This fact is exploited in many inference algorithms for discrete graphical models, but this is not true for continuous random variables. Characterizing $\mathcal{M}$ for general continuous sufficient statistics is in general very challenging. It is closely connected to the so-called *moment problem* [Landau, 1987] which has been studied since the late 19th century. For polynomial sufficient statistics, $\mathcal{M}$ can be characterized by a sequence of semidefinite constraints on the moments [Lasserre, 2009]. This is suggested by the positive semidefinite constraint on the covariance matrix Σ for Gaussian densities.

We now state several important facts relating the natural parameters θ and the mean parameters μ. For proofs, see [Wainwright and Jordan, 2008].

Lemma 2.3.1. Suppose θ corresponds to the natural parameters of an exponential family with sufficient statistics $\phi(x)$ and corresponding mean vector $\mu(\theta) = \mathbb{E}_{p_\theta}[\phi]$. Let $Z(\theta)$ be the corresponding log-partition function, and define its gradient $\nabla Z(\theta) : \mathcal{P} \to \mathcal{M}$. The following hold:

1. θ and μ are related by the mapping

$$\nabla Z(\theta) = \mu(\theta); \tag{2.7}$$

2. $\nabla^2 Z(\theta) = \text{cov}_\theta[\phi]$;

3. $Z(\theta)$ is a convex function, and strictly so if the family is minimal, so that $\delta^\top(\text{cov}_\theta[\phi])\delta > 0$ for each $\delta \neq 0$;

4. The mapping $\nabla Z(\theta) : \mathcal{P} \to \mathcal{M}$ is one-to-one if and only if the family is minimal;

14

5. The mapping $\nabla Z(\theta)$ is onto the interior of $\mathcal{M}$, $\mathrm{int}(\mathcal{M})$ if the family is minimal.

Remark 2.3.2. The exponential series family is minimal when the orthogonal series satisfies the *Haar condition*, so its likelihood is strictly convex, and (2.3) is a convex problem so long as the regularizer $\mathcal{R}$ is convex.

2.4 Duality

For any function Z taking values $\theta \in \mathcal{P}$, we define the *Fenchel conjugate*, Z^*, as follows:

$$Z^*(\mu) = \sup_{\theta \in \mathcal{P}} \left\{ \langle \theta, \mu \rangle - Z(\theta) \right\}, \tag{2.8}$$

When Z corresponds to the log-partition function this equation bears a strong resemblance to the maximum of the log-likelihood (2.3). Indeed when μ corresponds to empirical mean parameters $\widehat{\mu} = \frac{1}{n} \sum_{k=1}^{n} \phi(X^k)$ it is precisely that, though (2.8) is well-defined when μ doesn't correspond to a $\mu \in \mathcal{M}$.

For $\mu \in \mathrm{int}\mathcal{M}$ corresponding to a minimal family, let $\theta(\mu)$ denote the unique natural parameters corresponding to μ. Denote

$$H(\mu) := H(p_{\theta(\mu)}) = -\mathbb{E}_{p_{\theta(\mu)}}[\log p_{\theta(\mu)}] \tag{2.9}$$

the entropy of the density $p_{\theta(\mu)}$. Furthermore, denote the univariate entropy by

$$H_i(\mu) := -\mathbb{E}_{p_{\theta(\mu)}}\left[\log p_{i,\theta(\mu)}(X_i) \right], \tag{2.10}$$

and bivariate mutual information

$$I_{ij}(\mu) := \mathbb{E}_{p_{\theta(\mu)}}\left[\log \left(\frac{p_{ij\theta(\mu)}(X_i, X_j)}{p_{i,\theta(\mu)}(X_i)p_{j,\theta(\mu)}(X_j)} \right) \right]. \tag{2.11}$$

Theorem 2.4.1. *The Fenchel conjugate of the log-partition function Z is given by*

$$Z^*(\mu) = \begin{cases} -H(\mu), & \mu \in int\mathcal{M}; \\ \\ \infty, & \mu \notin \mathcal{M}; \end{cases} \tag{2.12}$$

for $\mu \in \mathcal{M} \backslash int\mathcal{M}$, $Z^*(\mu)$ is given by the limit of $Z^*(\mu^k)$ for any sequence $\{\mu^k\}$, $\mu^k \in int\mathcal{M}$.

Example 2.4.1 (Gaussian Entropy). Let $X \sim N(0, \Sigma)$, $\Sigma \succ 0$, and $\Omega = \Sigma^{-1}$. Denote p the density of X. Then

$$-\mathbb{E}_p[\log p] = \frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Omega| + \frac{1}{2} \mathbb{E}_p[X^\top \Omega X]. \tag{2.13}$$

Now, $X^\top \Omega X = \text{trace}(\Omega X X^\top)$, and since the trace is a linear operator, we may move the expectation inside the trace, giving

$$\mathbb{E}_p[X^\top \Omega X] = \text{trace}\left(\Omega \mathbb{E}_p[X X^\top]\right) \tag{2.14}$$

$$= \text{trace}\left(\Omega \Sigma\right) \tag{2.15}$$

$$= \text{trace}(I_d) = d. \tag{2.16}$$

Thus the Gaussian entropy is

$$\frac{d}{2}\left(1 + \log(2\pi)\right) - \frac{1}{2} \log |\Omega|. \tag{2.17}$$

Example 2.4.2 (Tree Entropy and Maximum Likelihood Trees). Let θ be the parameters of a minimal exponential family which is tree-structured: that is, any edge parameters $\theta_{ij} = 0$ for $(i, j) \notin T$, where T is the edge set for a tree. Because of the tree density factorization

(1.4),

$$H(\mu(\theta)) = -\mathbb{E}[\log p_{\theta(\mu)}] \tag{2.18}$$

$$= \sum_{i \in V} H_i(\mu(\theta)) - \sum_{(i,j) \in T} I_{ij}(\mu(\theta)). \tag{2.19}$$

Thus, for a tree-factored distribution has a simple expression for its entropy in terms of the univariate entropies and bivariate mutual informations.

2.5 Main Results

2.5.1 Sparsity

For our risk analysis, make the sparsity assumption on θ^*, that $\theta^* \in \tilde{\mathcal{P}}$, where

$$\tilde{\mathcal{P}} = \tilde{\mathcal{P}}(E) := \left\{ \theta : \|\theta_{ij}\|_2 = 0, \forall (i,j) \notin E \right\} \subseteq \mathcal{P}. \tag{2.20}$$

However, the set E is unknown. Recall that for the pairwise graphical model, $(i,j) \notin E$ when $\theta_{kl}^{ij} = 0$ for each $k, l = 1, \dots, \infty$; in other words, when $\|\theta^{i,j}\|_2 = 0$. For clarity, we refer to θ_v^*, μ_v to be the vectors of vertex parameters, θ_e, μ_e to be the vectors of edge parameters, and θ_{ij}, μ_{ij} to be the vector of parameters corresponding to edge (i,j). To encourage edge sparsity we consider the penalty

$$\mathcal{R}(\theta_e) = \sum_{i,j \in V, i < j} \|\theta_{ij}\|_2. \tag{2.21}$$

$\mathcal{R}$ defines a norm over the edge parameters. Furthermore, $\mathcal{R}$ has some special properties which we detail below.

Proposition 2.5.1. *The dual norm of $\mathcal{R}$ is*

$$\mathcal{R}^*(\theta_e) = \max_{(i,j)\in E} \|\theta_{ij}\|_2. \tag{2.22}$$

$\mathcal{R}$ is known as the (1,2)-group penalty, and its dual the $(\infty, 2)$-group penalty. In our application, the groups correspond to parameters of given edges. Group penalties are best known from their use in the group lasso [Yuan and Lin, 2006], which is used to encourage group sparsity in regression coefficients. For a vector of parameters θ denote its projection onto $\tilde{\mathcal{P}}$ by $\theta_{\tilde{\mathcal{P}}} = \sum_{(i,j)\in E} \|\theta_{ij}\|$, and its projection onto its orthogonal complement by $\theta_{\tilde{\mathcal{P}}^\perp} = \sum_{(i,j)\in E^c} \|\theta_{ij}\|$.

Proposition 2.5.2. $\mathcal{R}$ *is decomposable with respect to* $\tilde{\mathcal{P}}$. *That is,*

$$\mathcal{R}(\delta + \theta) = \mathcal{R}(\delta) + \mathcal{R}(\theta), \tag{2.23}$$

for each $\delta \in \tilde{\mathcal{P}}^\perp$ *and* $\theta \in \mathcal{P}$.

The following proposition characterizes the *subspace compatibility constant* for $\mathcal{R}$, which is necessary in the proofs.

Proposition 2.5.3.

$$\sup_{u\in\tilde{\mathcal{P}}\setminus\{0\}} \frac{\mathcal{R}(u)}{\|u\|} \leq \sqrt{|E|}. \tag{2.24}$$

We will state our main theoretical result for the regularized exponential series MLE. A full derivation of the results are in Appendix A.

We start with three assumptions:

Assumption 2.5.4. *Haar Condition*: Any truncated collection of basis elements$\{\bar{\phi}\}$ is linearly independent $\nu - a.e..$

Assumption 2.5.5. The univariate basis functions satisfy for each k, $|\phi_k| \le b(k) = O(k^\alpha)$ for some $\alpha \ge 0$.

Assumption 2.5.6. For each $i, j \in V$,

$$\underline{\epsilon} \le p_{ij}(x_i, x_j) \le \bar{\epsilon}, \tag{2.25}$$

for absolute constants $\underline{\epsilon} > 0, \bar{\epsilon} < \infty$.

These assumptions are mild. Many bases satisfy assumptions (1) and (2), such as the standard polynomial or trigonometric bases. For the orthonormal Legendre basis, $|\phi_k| \le \sqrt{2k+1}$, so it satisfies Assumption 2.5.5 with $\alpha = \frac{1}{2}$. The use of an overcomplete basis creates some statistical difficulties as the resulting truncated exponential family is no longer minimal. Assumption 2.5.6 is mild for density estimation as we only require boundedness of the bivariate marginals of p rather than of p itself.

The natural first question is whether the problem (2.3) has a solution at all, and if so, how many solutions. We begin by showing the existence and uniqueness of (2.3).

Lemma 2.5.7. Suppose $n > m_1$. The solution (2.3) exists and is unique with probability one.

2.5.2 Risk Consistency

We now present consistency of $p_{\hat{\theta}}$ in terms of the KL risk.

Theorem 2.5.8. *Suppose that the regularization parameter is chosen to be*

$$\lambda_n \asymp \sqrt{\frac{m_2^{2+4\alpha} \log(m_2 d)}{n}}, \tag{2.26}$$

and the truncation parameters satisfy

$$m_1 = \Omega\left(d^{\frac{1}{r-\alpha-1/2}}\right), \tag{2.27}$$

$$m_2 = \Omega\left(|E|^{\frac{1}{r-\alpha-1/2}}\right). \tag{2.28}$$

Then the regularized exponential series MLE $\widehat{\theta}_\lambda$ satisfies

$$KL(p \mid p_{\widehat{\theta}}) = O_p\left(m_2^{-2r}|E| + m_1^{-2r}d + \frac{m_2^{2+4\alpha}\log(dm_2)}{n}|E| + \frac{m_1}{n}d\right). \tag{2.29}$$

Corollary 2.5.9. *The optimal choice of truncation dimensions m_1, m_2 is*

$$m_1 \asymp \max\left\{n^{\frac{1}{2r+1}}, d^{\frac{1}{r-\alpha-1/2}}\right\},$$

$$m_2 \asymp \max\left\{n^{\frac{1}{2r+2+4\alpha}}, |E|^{\frac{1}{r-\alpha-1/2}}\right\}.$$

Consider typical choices of $r = 2$, $\alpha = \frac{1}{2}$ (such as the Legendre basis).

- *The dimension d and edge cardinality $|E|$ may scale as*

$$d = o(\sqrt{n}), \tag{2.30}$$

$$|E| = o(n^{\frac{1}{5}}/\log(n)), \tag{2.31}$$

  with the risk still approaching zero as $n \to \infty$.

- *Suppose that $d = O(n^{\frac{1}{5}})$ and $|E| = O(n^{\frac{1}{8}})$. Then by choosing*

$$m_1 \asymp n^{\frac{1}{5}}, \tag{2.32}$$

$$m_2 \asymp n^{\frac{1}{8}}, \tag{2.33}$$

$$\lambda \asymp \frac{\log^{1/2}(nd)}{n^{\frac{1}{4}}}, \tag{2.34}$$

20

the risk decreases as

$$KL(p \mid p_{\widehat{\theta}}) = O_p \left(\frac{|E| \log(dn)}{n^{\frac{1}{2}}} + \frac{d}{n^{\frac{4}{5}}} \right). \tag{2.35}$$

Remark 2.5.10. The result of 2.5.8 holds uniformly for any set B of pairwise densities p with bounded Sobolev norm. In particular, assuming $\mid E \mid = o(n^{1/2})$ and $d = o(n^{4/5})$,

$$\lim_{t \to \infty} \lim_{n \to \infty} \sup_{p \in B} \mathbb{P} \left(KL(p \mid \widehat{p}) \geq \frac{\mid E \mid \log dn}{n^{\frac{1}{2}}} t + \frac{d}{n^{\frac{4}{5}}} t \right) = 0. \tag{2.36}$$

Remark 2.5.11. Theorem 2.5.8 shows that the risk of $p_{\widehat{\theta}}$ adapts to the unknown sparsity of θ^*, in that the risk contains a factor of $|E|$ rather than d^2. However, this is not sufficient for model selection consistency, which requires further assumptions. In particular consistency in KL risk does not require an *incoherence* condition. We consider model selection in the next section.

2.5.3   Model Selection

Our result for model selection consistency requires more stringent assumptions, in addition to those in the previous section. We denote the vector of truncated parameters by $\bar{\theta}^*$. We index the (infinite) vector of omitted parameters by T, so that the vector of parameters are θ_T^*. We use the subscript E to denote the collection of parameters of edges in E (as well as all vertex parameters), and E^c to denote parameters for edges in E^c. Denote the covariance matrix of the sufficient statistics ϕ by

$$\Gamma := \mathrm{cov}_p[\phi]. \tag{2.37}$$

For two index sets A, B denote Γ_{AB} to be the cross-covariance between ϕ_A and ϕ_B, $\mathrm{cov}[\phi_A, \phi_B]$.

Assumption 2.5.12. For a constant $\kappa_\Gamma < \infty$,

$$\|\Gamma_{EE}^{-1}\|_2 \leq \frac{\kappa_\Gamma}{m\sqrt{d+|E|}}, \tag{2.38}$$

$$\kappa_T := \|\Gamma_{ET}\|_\infty. \tag{2.39}$$

where $\|A\|_\infty = \max_i \sum_j |A_{ij}|$ here denotes the matrix ∞ norm and $\|A\|_2$ the matrix operator norm. Additionally, we define the following:

$$\bar{K}_{\theta^\dagger} := \mathbb{E}_{\theta^\dagger}\left[U \cdot (\bar\phi - \mathbb{E}_{\theta^\dagger}[\bar\phi]) \right], \tag{2.40}$$

where

$$U = (\|\bar\phi_E - \mathbb{E}_{\theta^\dagger}[\bar\phi_E]\|_1 + \|\phi_T - \mathbb{E}_{\theta^\dagger}[\phi_T]\|_1)^2, \tag{2.41}$$

and $\theta^\dagger := \theta^* + z(\widehat\theta - \theta^*)$ and $z \in [0,1]$.

Assumption 2.5.13. For some $\kappa_R < \infty$, for all $z \in [0,1]$ and all $\widehat\theta$ satisfying

$$\widehat\theta_{E^c} = 0, \tag{2.42}$$

$$\|\bar\theta_E^* - \widehat\theta_E\|_\infty \leq 2\kappa_\Gamma \left(\|\widehat\mu_E - \mu_E^*\|_\infty + \lambda_n/m + (\kappa_T + 1)\|\theta_T^*\|_\infty\right), \tag{2.43}$$

we have that

$$\|\bar{K}_{\theta^\dagger}\|_\infty \leq \kappa_R \max\{b(m_2)^2, b(m_1)\}. \tag{2.44}$$

This assumption may appear opaque so we will elaborate. If p_θ is bounded, the third central moment of the univariate statistic ϕ_k is

22

$$\mathbb{E}_\theta[(\phi_k - \mathbb{E}_\theta[\phi_k])^3] \leq b(k)\mathbb{E}_\theta[(\phi_k - \mathbb{E}_\theta[\phi_k])^2] \tag{2.45}$$

$$\leq \bar{\epsilon}b(k)\int \phi_k^2 \tag{2.46}$$

$$\leq \bar{\epsilon}b(k). \tag{2.47}$$

This holds similarly for bivariate sufficient statistics. $\bar{K}_{\theta\dagger}$ is a sum of a third central moment of a sufficient statistic and the third cross-moments of that statistic with the other sufficient statistics. We thus require that the third cross central moments between sufficient statistics decay sufficiently rapidly, so that the sum is on the order as stated. In our proof, this factors in to the *remainder term*, which is the bias from truncating the infinite expansion of the log density. Finally, we have an *irrepresentable condition*

Assumption 2.5.14.

$$\max_{(i,j)\in E^c} \|\Gamma_{ij,E}\Gamma_{EE}^{-1}\|_2 \leq \frac{1-\tau}{\sqrt{d+E}}, \qquad \text{for some } \tau \in (0,1]. \tag{2.48}$$

Here $\|A\|_2$ is the matrix operator norm. This condition is reminiscent of the irrepresentable condition for sparse additive models in [Ravikumar et al., 2009], in that it involves the operator norm rather than the matrix ∞ norm. It guarantees that no sets of variables in E and E^c are too strongly influenced. In the following theorem we assume $\kappa_\Gamma, \kappa_T, \kappa_R$ grow as constants, though they are tracked in the supplementary lemmas. Finally we define $\rho^* = \min_{(i,j)\in E} \|\bar{\theta}^*\|_\infty$ to be the minimum ∞ norm of the edge parameters.

Theorem 2.5.15. *Denote $\widehat{E}$ to be the edge set learned from $\widehat{\theta}_{\lambda_n}$; $\widehat{E} := \{(i,j) : \|\widehat{\theta}_{ij}\|_2 = 0\}$.*

If the truncation dimensions m_1, m_2 and regularization parameter λ_n satisfy

$$m_2 \asymp n^{\frac{1}{2r+4\alpha+1}}, \tag{2.49}$$

$$m_1 \asymp n^{\frac{1}{2r+4\alpha+1}}, \tag{2.50}$$

$$\lambda_n \asymp \sqrt{\frac{\log(nd)}{n^{\frac{2r-1}{2r+4\alpha+1}}}} \tag{2.51}$$

and suppose that the number of variables d and ρ^* satisfy

$$d = o\left(e^{n^{\frac{2r-1}{2r+4\alpha+1}}}\right), \tag{2.52}$$

$$\frac{1}{\rho^*} = o\left(\sqrt{\frac{n^{\frac{2r+1}{2r+1+4\alpha}}}{\log(nd)}}\right), \tag{2.53}$$

then

$$\mathbb{P}(\widehat{E} = E) \to 1. \tag{2.54}$$

2.6   Discussion

Proofs and supporting lemmas for this chapter may be found in Chapter 5, but we will briefly discuss the results here. In [Barron and Sheu, 1991], the optimal choice of truncation for univariate exponential series approximation was $m_1 \asymp n^{\frac{1}{2r+1}}$, and for the bivariate problem we have $m_2 \asymp n^{\frac{1}{2r+2}}$. Our truncation is of a lower order. In our proofs, in order for our estimator to adapt to the unknown sparsity of E we require exponential concentration for the sufficient statistics. To do this, we use Hoeffding's inequality, which gives that

$$\|\widehat{\mu}_{ij} - \mu_{ij}\| = O_p\left(\sqrt{\frac{m_2^{2+4\alpha}}{n}}\right). \tag{2.55}$$

[Barron and Sheu, 1991] use Chebyshev's inequality, gives a tighter bound of $\sqrt{\frac{m_2^2}{n}}$ but doesn't give exponential concentration.

Our model selection results require conditions on the covariance matrix of the sufficient statistics Γ, particularly an incoherence condition. This is natural; for example, for Gaussian graphical model selection, the same incoherence condition is required on the covariance matrix of the sufficient statistics; in this application the covariance has the simple expression $\Sigma \otimes \Sigma$ [Ravikumar et al., 2011]. For the typical choices of $r = 2$ and $\alpha = \frac{1}{2}$, the dimension may scale nearly exponentially with the sample size,

$$d = o\left(e^{n^{\frac{3}{7}}}\right), \tag{2.56}$$

and ρ^* may scale as

$$\frac{1}{\rho^*} = o\left(\frac{n^{5/14}}{\log^{1/2}(nd)}\right), \tag{2.57}$$

with $\widehat{E} = E$ with probability approaching one. Observe that the KL risk of $p_{\widehat{\theta}}$ may diverge rapidly while still having the correct sparsity pattern with high probability. For the parametric Gaussian graphical model the optimal rate is $o(e^n)$, [Ravikumar et al., 2011] which is also the optimal rate for forests [Liu et al., 2012b]. The optimal choice of m_2 for model selection is larger than that for the risk analysis, so by *oversmoothing* the edge potentials, we get better sample complexity for the model selection problem. This phenomenon was also found for graph selection for forests [Liu et al., 2012b].

CHAPTER 3

TREE-REWEIGHTED VARIATIONAL LIKELIHOOD APPROXIMATION

First-order optimization procedures for solving the regularized maximum likelihood problem requires evaluation of $Z(\theta)$ and its gradient $\nabla Z(\theta) = \int_{\mathcal{X}} p_\theta(x)\phi(x)d\nu(x)$. For a pairwise density on d nodes, these computations still require a d-dimensional integral even when the graph is sparse. Using the junction tree algorithm [Koller and Friedman, 2009], it is possible to factorize the joint density into terms which have no more variables than the *treewidth* of the graph, making these calculations simpler. However, the treewidth of a graph may in general be large, and we are interested in procedures which work for general graphs.

Monte Carlo methods are one popular approach for approximating partition functions [Gilbert and Nocedal, 1992]. However, it may take a very long time for suitable convergence, and such methods generally don't provide finite-time bounds on the accuracy of the approximation.

Here we pursue a tree-reweighted variational approach [Wainwright et al., 2005], which replaces $Z(\theta)$ by a surrogate $Q(\theta)$. Our method has several key advantages:

1. It computes an approximation to $Z(\theta)$ and its gradient $\nabla Z(\theta)$ all in one pass using a parallelizable message passing algorithm;

2. $Q(\theta)$ is guaranteed to be an upper bound on $Z(\theta)$ for each θ, with the tightness dictated by variational parameters;

3. It is based on convex optimization of the variational parameters, so there are deterministic stopping criteria for computing $Q(\theta)$;

4. $Q(\theta)$, like $Z(\theta)$, will be strictly convex in θ, so there are criterion for global convergence of the approximate maximum likelihood.

3.1 Problem Formulation

Recall that a density with a tree graph can be factorized in the form (1.4). A density corresponding to a graph G with cycles cannot in general be factorized. Instead we will consider the collection $\mathcal{T}$ of spanning trees of G. Consider an exponential family following the graph G having natural parameters θ. For a spanning tree $T \in \mathcal{T}$, consider the vector θ^T that obeys T: $\theta_{ij}^T = 0$ if $(i,j) \notin T$; in shorthand we write θ_{ij}^T to be the vector of parameters corresponding to edge (i,j).

Now consider writing the parameter value θ as a convex combination of spanning tree parameters:

$$\theta = \sum_{T \in \mathcal{T}} \alpha_T \theta^T,$$

by the convexity of the log-partition function $Z(\theta)$, we have

$$Z(\theta) \leq \sum_{T \in \mathcal{T}} \alpha_T Z\left(\theta^T\right). \tag{3.1}$$

Now, we may form the tightest upper bound on the log-partition function by solving

$$Q(\theta, \alpha) := \min_{\{\theta^T\}_{T \in \mathcal{T}}} \left\{ \sum_{T \in \mathcal{T}} \alpha_T Z\left(\theta^T\right) \right\} \tag{3.2}$$

$$s.t. \quad \theta = \sum_{T} \alpha_T \theta^T.$$

Observe that since Z is a convex function, (3.2) is convex in $\{\theta^T\}_{T \in \mathcal{T}}$, and the constraints are linear, so the problem is convex. However, the number of spanning trees of a general loopy graph G could be very large, perhaps even super-exponential in the number of edges [Cayley, 1889], so the number of parameters to minimize over is in general very large. Contrary to expectation, it is possible to efficiently solve this problem. To begin, we will look at the dual

problem to (3.2).

3.2   Solution to the Dual Problem

For an edge (i, j), write

$$\alpha_{ij} = \sum_{T \in \mathcal{T}} \alpha_T 1\{(i, j) \in T\}. \tag{3.3}$$

This is the *edge appearance probability*: the probability edge (i, j) is observed when drawing a spanning tree at random according to the distribution $\{\alpha_T\}$. The set of such edge appearance probability vectors $\{\alpha_{ij}\}$ which can be written as convex combinations of tree indicator vectors, is known as the *spanning tree polytope*, which we denote $\mathcal{S}_{\mathcal{T}}$. Let

$$\tilde{\Pi} = \left\{ \{q_i, q_{ij}\} : \int_{\mathcal{X}_j} q_{ij}(x_i, x_j) dx_j = q_i(x_i), \int_{\mathcal{X}_i \times \mathcal{X}_j} q_{ij} = 1, \tag{3.4} \right.$$
$$\left. \int_{\mathcal{X}_i} q_i = 1, q_{ij} \geq 0, q_i \geq 0, (i, j) \in E \right\},$$

$$\tilde{\mathcal{M}} = \left\{ \mu : \exists \{q_i, q_{ij}\} \in \tilde{\Pi} : \mathbb{E}_{q_i}[\phi_i] = \mu_i, \mathbb{E}_{q_{ij}}[\phi_{ij}] = \mu_{ij}, (i, j) \in E \right\}. \tag{3.5}$$

That is, $\tilde{\Pi}$ is the set of univariate and bivariate densities over E which respect marginalization, and $\tilde{\mathcal{M}}$ is the set of mean parameters which can arise from elements of $\tilde{\Pi}$. To derive the dual problem to (3.2), we first define the Lagrangian

$$L(\{\theta^T\}_{T \in \mathcal{T}}, \tau) = \sum_{T \in \mathcal{T}} \alpha_T Z\left(\theta^T\right) + \tau^\top \left(\theta - \sum_{T \in \mathcal{T}} \alpha_T \theta^T\right) \tag{3.6}$$

$$= \langle \tau, \theta \rangle - \sum_{T \in \mathcal{T}} \alpha_T \left(\langle \tau, \theta^T \rangle - Z(\theta^T)\right) \tag{3.7}$$

To minimize L with respect to θ^T for a given $T \in \mathcal{T}$, we set the associated derivative to zero: $\frac{\partial L}{\partial \theta^T} = 0$. Denoting the optimum by θ_*^T, the solution may be written in terms of the

Fenchel conjugate of Z (section 2.4):

$$Z_T^*(\tau) := \max_{\theta^T} \left\{ \langle \tau, \theta^T \rangle - Z(\theta^T) \right\} = \langle \tau, \theta_*^T \rangle - Z(\theta_*^T). \tag{3.8}$$

Recall that the dual of Z for a tree-structured parametrization (2.19) takes the form

$$Z_T^*(\tau) = \sum_{i \in V} H_i(\tau) - \sum_{(i,j) \in T} I_{ij}(\tau), \qquad \tau \in \tilde{\mathcal{M}}, \tag{3.9}$$

and so the dual to (3.2) is

$$\max_{\tau \in \tilde{\mathcal{M}}} \left\{ \langle \theta, \tau \rangle + \sum_{i \in V} H_i(\tau) - \sum_{(i,j) \in E} \alpha_{ij} I_{ij}(\tau) \right\}. \tag{3.10}$$

For some distributions, such as discrete pairwise models and Gaussian models, the entropy and mutual information have a closed form and (3.10) can be solved explicitly. Unfortunately, for continuous models there is typically no such expression. In the following section we show that (3.10) is equivalent to a functional optimization problem, which we solve using message passing.

3.3   Functional Message Passing

Observe that the dual problem finds an optimum over the space of mean parameters realizable by distributions in $\tilde{\Pi}$, so (3.10) is equivalent to the following functional optimization:

$$\max_{q \in \tilde{\Pi}} \left\{ \sum_{i \in V} \left(\langle \theta_i, \mathbb{E}_{q_i}[\phi_i] \rangle - \mathbb{E}_{q_i}[\log q_i] \right) \right.$$
$$\left. + \sum_{(i,j) \in E} \left(\langle \theta_{ij}, \mathbb{E}_{q_{ij}}[\phi_{ij}] \rangle - \alpha_{ij} \mathbb{E}_{q_{ij}} \left[\log \left(\frac{q_{ij}}{q_i q_j} \right) \right] \right) \right\} \tag{3.11}$$

If q_i^*, q_{ij}^* are solutions to (3.11), the solution to the dual problem (3.10) is given by

$$\tau_i^*(\theta) = \mathbb{E}_{q_i^*}[\phi_i], \qquad\qquad i \in V, \qquad\qquad (3.12)$$

$$\tau_{ij}^*(\theta) = \mathbb{E}_{q_{ij}^*}[\phi_{ij}], \qquad\qquad (i,j) \in E. \qquad\qquad (3.13)$$

The optimization (3.11) is an optimization of a convex functional over a space of linear functional constraints $\tilde{\Pi}$. Any solution to the stationary conditions of the associated Lagrangian will thus correspond to a global optimum. We may derive the stationary conditions using standard arguments from calculus of variations. Write the constraints as

$$C_i(q_i) = 1 - \int q_i(x_i)\, dx_i \qquad\qquad (3.14)$$

$$C_{ij}(x_j, q_{ij}) = q_j(x_j) - \int q_{ij}(x_i, x_j)\, dx_i, \qquad\qquad (3.15)$$

$$C_{ji}((x_i, q_{ij}) = q_i(x_i) - \int q_{ij}(x_i, x_j)\, dx_j, \qquad\qquad (3.16)$$

and let $\eta_i, \eta_{ij}(x_j), \eta_{ji}(x_i)$ be the Lagrange multipliers associated with these constraints. The second and third multipliers are real-valued functions. This gives us the stationary conditions

$$\log q_i(x_i) = \langle \theta_i, \phi_i(x_i) \rangle + \sum_{r \in N(i)} \eta_{ri}(x_i) + \eta_i, \qquad\qquad (3.17)$$

$$\alpha_{ij} \log \frac{q_{ij}(x_i, x_j)}{q_i(x_i) q_j(x_j)}) = \langle \theta_{ij}, \phi_{ij}(x_i, x_j) \rangle - \eta_{ji}(x_i) - \eta_{ij}(x_j), \qquad\qquad (3.18)$$

we may simplify the second condition to get

$$\begin{aligned}
\log q_{ij}(x_i, x_j) = {}& \eta_i + \eta_j + \langle \theta_{ij}, \phi_{ij}(x_i, x_j) \rangle / \alpha_{ij} \\
& + \langle \theta_i, \phi_i(x_i) \rangle + \langle \theta_j, \phi_j(x_j) \rangle \\
& + \sum_{r \in N(i) \backslash j} \eta_{ri}(x_i) / \alpha_{ij} + \sum_{r \in N(j) \backslash i} \eta_{rj}(x_j) / \alpha_{ij}.
\end{aligned} \qquad (3.19)$$
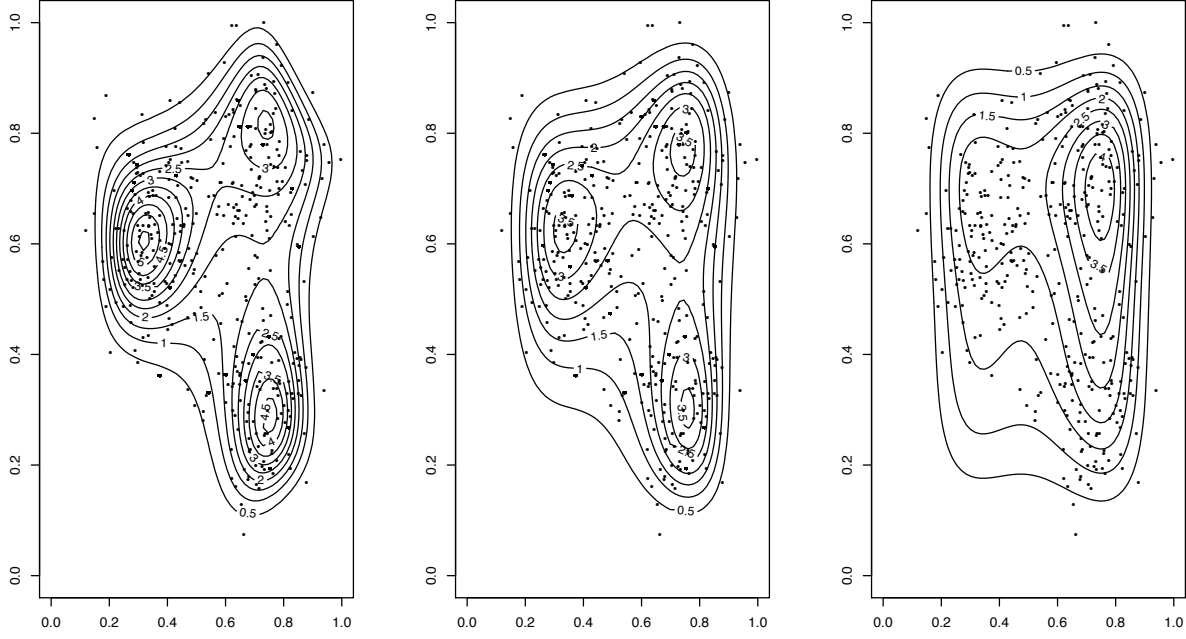
Figure 3.1: Bivariate pseudo densities under increasing regularization. Left plot is unregularized; right plot is fully regularized (dimensions are independent). Simulated data is a mixture of three spherical Gaussians.

For each $(i, j) \in E$, we define a message $M_{ij} : \mathcal{X}_j \to \mathbb{R}$, and $M_{ji} : \mathcal{X}_i \to \mathbb{R}$, by

$$M_{ij}\left(x_j\right) \;=\; e^{\eta_{ij}\left(x_j\right)/\alpha_{ij}}, \tag{3.20}$$

$$M_{ji}\left(x_i\right) \;=\; e^{\eta_{ji}\left(x_i\right)/\alpha_{ij}}, \tag{3.21}$$

so that the solution to (3.11) takes the form

$$q_i\left(x_i\right) \;\propto\; \exp\{\langle\theta_i, \phi_i\rangle\} \prod_{j \in N(i)} M_{ji}\left(x_i\right)^{\alpha_{ji}}, \tag{3.22}$$

$$q_{ij}\left(x_i, x_j\right) \;\propto\; \exp\{\langle\theta_{ij}, \phi_{ij}\rangle/\alpha_{ij} + \langle\theta_i, \phi_i\rangle + \langle\theta_j, \phi_j\rangle\} \tag{3.23}$$

$$\times \frac{\prod_{r \in N(i)\backslash j} M_{ri}\left(x_i\right)^{\alpha_{ri}} \prod_{r \in N(j)\backslash i} M_{rj}\left(x_j\right)^{\alpha_{rj}}}{M_{ji}\left(x_i\right)^{1-\alpha_{ji}} M_{ij}\left(x_j\right)^{1-\alpha_{ij}}}.$$

These are *pseudodensities*: they are valid densities which obey the marginalization con-

31

1. Initialize messages $\left\{ M_{ij}^0, M_{ji}^0 \right\}$;

2. For $n = 0, 1, \ldots$:
   Update beliefs:
   $$b_i^{n+1}(x_i) \propto \exp \langle \theta_i, \phi_i \rangle \prod_{r \in N(i)} M_{ri}^n (x_i)^{\alpha_{ri}} ; \qquad (3.25)$$

   Update messages:

   $$M_{ij}^{n+1}(x_j) \leftarrow \int_{\mathcal{X}_i} \exp\{\langle \theta_{ij}, \phi_{ij} \rangle / \alpha_{ij}\} \left\{ \frac{b_i(x_i)}{M_{ji}^n(x_i)} \right\} dx_i. \qquad (3.26)$$

Figure 3.2: Functional Belief Propagation

straints, but they may not together correspond to the marginal distributions of any higher-dimensional joint distribution. By enforcing the marginalization constraints for $\{q_i\}$, $\{q_{ij}\}$, we find that the messages follow the fixed-point conditions

$$
\begin{aligned}
M_{ij}(x_j) &\propto \int_{\mathcal{X}_i} \exp\{\langle \theta_{ij}, \phi_{ij} \rangle / \alpha_{ij} + \langle \theta_i, \phi_i \rangle\} \frac{\prod_{r \in N(i) \setminus j} M_{ri}(x_i)^{\alpha_{ri}}}{M_{ji}(x_i)^{1-\alpha_{ji}}} dx_i \\
&= \int_{\mathcal{X}_i} \exp\{\langle \theta_{ij}, \phi_{ij} \rangle / \alpha_{ij}\} \left\{ \frac{q_i(x_i)}{M_{ji}(x_i)} \right\} dx_i.
\end{aligned} \qquad (3.24)
$$

To find the fixed point corresponding to the pseudomarginal densities we run the algorithm in figure 3.3 to convergence:

The beliefs are normalized to integrate to 1, so they correspond to a proper density. At convergence, the beliefs $\{b_i\}$ correspond to the univariate pseudodensities $\{q_i^*\}$. The message updates can be performed in parallel. Furthermore, more elaborate schedules exist which may speed up convergence. For example, updates can be formed dynamically. Also, message updates corresponding to beliefs which have reached convergence can be skipped. See [Gonzalez et al., 2011] for a treatment on different parallel scheduling methods. If the fixed point updates do converge, they will converge to the unique fixed point corresponding to the global minimum of (3.10). The fixed-point updates are not guaranteed to converge.

In our experiments we only encountered stability issues after taking too large of a step in the ISTA algorithm for estimation, resulting in an unstable candidate step; if we encounter a convergence problem we simply take a smaller step. As such we don't find the need for damping or other techniques to encourage convergence.

To perform message passing in practice we discretized messages and approximated integrals using a Riemann sum approximation. We found this to give very accurate results in experiments. Other approximations for continuous message passing exist [Noorshams and Wainwright, 2013; Sudderth et al., 2010] which could be more memory and computation efficient for large problems.

3.4   Optimizing Edge Weights

The previous analysis outlines how to compute $Q(\theta, \alpha)$ for a set of fixed edge weights $\{\alpha_{ij}\}$. In this section we show how the edge weights can be optimized to produce tighter bounds on the likelihood by solving

$$Q(\theta) := \min_{\alpha \in \mathcal{S}_{\mathcal{T}}} Q(\theta, \alpha). \tag{3.27}$$

Our analysis follows that of [Wainwright et al., 2005]. From Danskin's theorem [Bertsekas, 1999], observing the form of (3.10) it follows that the function $Q(\theta, \alpha)$ is convex as a function of α, with gradient

$$\nabla_{\alpha_{ij}} Q(\theta, \alpha) = -I_{ij}(\tau^*(\theta)), \tag{3.28}$$

where $\tau^*(\theta)$ are the pseudomoments from solving the dual problem (3.12). (3.27) is the minimization of a convex objective over a convex polytope, so it is a convex problem. However, the number of constraints characterizing $\mathcal{S}_{\mathcal{T}}$ are typically prohibitively large. To avoid dealing with them directly, we employ the following strategy. Suppose we have the
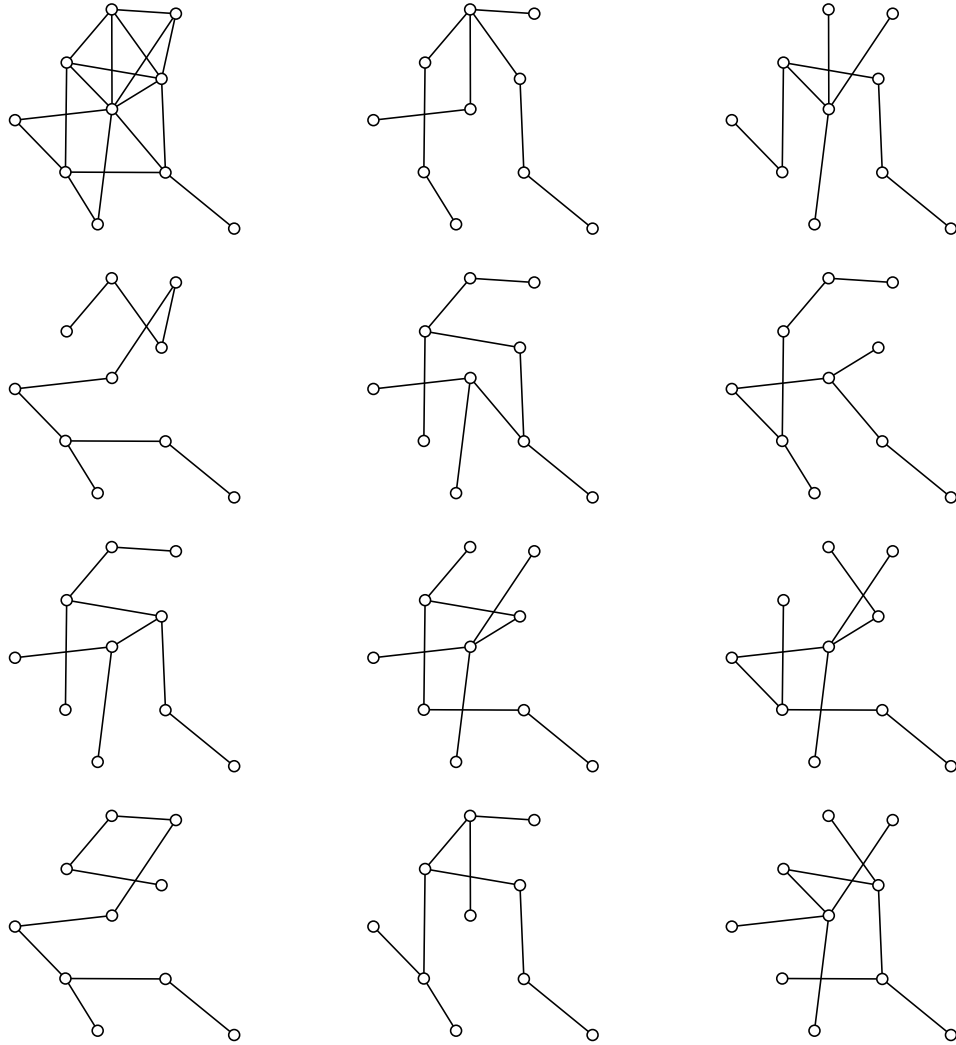
Figure 3.3: Top left: A graph on 10 nodes. Rest: Corresponding spanning trees (not all figured).

Figure 3.4: Kruskal's Algorithm

current iterate α^t. We linearize $Q(\theta, \alpha)$ about α^t and solve

$$\min_s \quad \nabla_\alpha Q(\theta, \alpha)^\top (s - \alpha^t), \qquad (3.29)$$

$$s.t. \quad s \in \mathcal{S}_\mathcal{T}.$$

This is a linear program, so the solution must always fall on at least one vertex of $\mathcal{S}_\mathcal{T}$. From observing the structure of $\mathcal{S}_\mathcal{T}$ as being supported by spanning tree indicator vectors, a solution s is equal to the indicator vector of a maximum weight spanning tree with weights $\{I_{ij}(\tau^*(\theta^t))\}$. Finding the maximum weight spanning tree can be done efficiently in $O(|E| \log |E|)$ time using Kruskal's algorithm (Figure 3.4). Lastly we update the edge weights $\alpha^{t+1} \leftarrow c\alpha^t + (1 - c)s$, where c is a step size $c \in (0, 1)$. To ensure convergence guarantees, c can either be set to $c = \frac{2}{2+t}$, or it can be chosen using line search. This technique is known as the *Frank-Wolfe* algorithm [Bertsekas, 1999], and is known to converge at the rate of $O(1/t)$.

3.5 Variational Maximum Likelihood

Variational regularized maximum likelihood replaces the regularized maximum likelihood equation (2.3) with

$$\max_{\theta \in \mathcal{P}} \left\{ \langle \theta, \widehat{\mu} \rangle - Q(\theta) - \lambda \mathcal{R}(\theta_e) \right\}, \qquad (3.30)$$

$Q(\theta)$ is the variational approximation to the true log-partion function $Z(\theta)$.

This shares many features in common with the regularized MLE. For example, by Danskin's theorem $\nabla Q(\theta) = \tau^*(\theta)$, the optimal pseudomoments from message passing. Furthermore, $\nabla^2_{a,b} Q(\theta) = \text{cov}_{q^*}(\phi_a, \phi_b)$, and the minimality of ϕ implies Q is strictly convex, since for any $a \neq 0$, $a^\top \nabla^2 Q(\theta) a = \text{var}_{q^*}(a^\top \phi) \neq 0 \ \nu - a.e..$

3.5.1 Optimization Algorithms

Both the exact optimization problem in (2.3) and the approximate problem (3.30) can be written as minimization of a smooth (strictly convex) function plus a non-smooth convex function,

$$\min_{\theta \in \mathcal{P}} \left\{ -\mathcal{L}(\theta) + \lambda \mathcal{R}(\theta) \right\}. \tag{3.31}$$

Several algorithms have been designed to solve problems of this form; for a review see [Bach et al., 2011]. We will focus on what are known as *proximal gradient* methods [Nesterov, 2013; Beck and Teboulle, 2009], a class of first-order methods which have proven effective for large scale, non-smooth optimization.

ISTA

The simplest such algorithm is called the *iterative-shrinkage thresholding algorithm*, or ISTA, which works as follows. Fix the current estimate θ^t. Linearize $\mathcal{L}$ about the current point and solve:

$$p_L(\theta^t) = \operatorname*{argmin}_{\theta \in \mathcal{P}} \left\{ \langle \nabla \mathcal{L}(\theta^t), \theta^t - \theta \rangle + \lambda \mathcal{R}(\theta) + \frac{L}{2} \|\theta - \theta^t\|^2 \right\}, \tag{3.32}$$

where L is a step size. The squared norm term is called the proximal term, which encourages the solution not to be too far from the current step θ^t. After some manipulation, is can be made equivalent to

$$p_L(\theta^t) = \operatorname*{argmin}_{\theta \in \mathcal{P}} \left\{ \frac{1}{2} \left\| \theta - \left(\theta^t - \frac{1}{L}(-\nabla \mathcal{L}(\theta^t)) \right) \right\|^2 + \frac{\lambda}{L} \mathcal{R}(\theta) \right\}. \tag{3.33}$$

1. Input current iterate θ^t;

2. Fix a $L_0 > 0$, $\delta > 1$;

3. Find the smallest nonnegative integer l such that for $L' = \delta^l L_0$,

$$\mathcal{L}(\theta^t) - \mathcal{L}(p_{L'}(\theta^t)) \leq \langle p_{L'}(\theta^t) - \theta^t, \nabla\mathcal{L}(\theta^t) \rangle + \frac{L'}{2}\|\theta^t - p_{L'}(\theta^t)\|^2, \qquad (3.36)$$

4. Update θ^{t+1} with (3.33), using $L = \delta^l L_0$.

Figure 3.5: ISTA Line Search

For the group regularizer $\mathcal{R}(\theta) = \sum_{(i,j)\in E} \|\theta_{ij}\|$, the solution has a closed form and is given by

$$(p_L(\theta^t))_i = \theta_i^t - \frac{1}{L}\left(-\nabla_i\mathcal{L}(\theta^t)\right); \qquad (3.34)$$

$$(p_L(\theta^t))_{ij} = \left(1 - \frac{\lambda/L}{\|\theta_{ij}^t - \frac{1}{L}(-\nabla_{ij}\mathcal{L}(\theta^t))\|}\right)_+ \left(\theta_{ij}^t - \frac{1}{L}(-\nabla_{ij}\mathcal{L}(\theta^t))\right), \qquad (3.35)$$

and we set the update steps to $\theta_i^t = (p_L(\theta^t))_i$ and $\theta_{ij}^t = (p_L(\theta^t))_{ij}$ for each $i \in V$ and $(i', j) \in E$. $(\cdot)_+ := \max(\cdot, 0)$. In the absence of regularization, the proximal gradient method simplifies to gradient descent. When $\lambda > 0$, due to the soft thresholding, it may produce exactly sparse solutions, where all edge parameters θ_{ij} for a particular edge (i, j) are zero. This is an advantage over other methods which only produce a sparse estimate up to numerical error, and necessitate truncation.

Since exact bounds on the Hessian of $\mathcal{L}$ aren't known, we must choose L using line search. We employ the backtracking line search from [Beck and Teboulle, 2009], in Figure 3.5.

FISTA

The accelerated counterpart to ISTA is the *fast iterative thresholding-scaling algorithm*, or FISTA [Beck and Teboulle, 2009]. It is analogous to the accelerated gradient method in

1. Input iterates $\theta^t, \theta^{t-1}, y^t$;

2. Fix a $L_0 > 0$, $\delta > 1$;

3. Find the smallest nonnegative integer l such that for $L' = \delta^l L_0$,

$$\mathcal{L}(y^t) - \mathcal{L}(p_{L'}(y^t)) \leq \langle p_{L'}(y^t) - y^t, \nabla \mathcal{L}(y^t) \rangle + \frac{L'}{2}\|y^t - p_{L'}(y^t)\|^2, \qquad (3.37)$$

4. Set $a_{t+1} = \frac{1+\sqrt{1+4a_t^2}}{2}$,

5. Update $\theta^t = p_L(y^t)$, using $L = \delta^l L_0$,

6. Update $y^{t+1} = \theta^t + \frac{a_t - 1}{a_{t+1}}(\theta^t - \theta^{t-1})$.

Figure 3.6: FISTA Line Search

smooth optimization, which has shown to be an optimal first-order method for smooth optimization [Nemirovsky and Yudin, 1983]. Instead of the new parameter iterates being a projection of the previous, it is a projection of a linear combination of the two previous iterates. The updates with line search are given in 3.6.

FISTA requires essentially the same computation at each iteration, in particular the same number of gradient evaluations. In addition to the standard proximal gradient algorithms, we found success initializing L_0 using a secant rule:

$$L_0 = \frac{\langle \theta^t - \theta^{t-1}, -\nabla \mathcal{L}(\theta^t) + \nabla \mathcal{L}(\theta^{t-1}) \rangle}{\|\theta^t - \theta^{t-1}\|^2}. \qquad (3.38)$$

Typically we find initializing with the secant rule finds a direction of sufficient descent with little backtracking.

Discussion

ISTA and its accelerated counterpart FISTA [Beck and Teboulle, 2009] have linear convergence, that is at the rate $O(C^t)$ for some $C < 1$, when $-\mathcal{L}$ is strongly convex. In contrast, if

$-\mathcal{L}$ is only Lipschitz, ISTA converges at the sub-linear rate $O(1/k)$, and FISTA at the rate $O(1/k^2)$. Recall that with strong convexity, gradient descent also has linear convergence, so proximal gradient descent behaves like gradient descent despite the objective not being smooth. Thus proximal gradient methods have superior theoretical guarantees to competitors such as subgradient descent. The negative log-likelihood $-\mathcal{L}$ is only strictly convex, but it is strongly convex in a neighborhood of the solution [Kakade et al., 2010]. Thus we may think of these proximal gradient methods converging linearly after a sufficient "burn-in" phase.

FISTA, like the accelerated gradient algorithm has been shown to outperform ISTA in some real problems [Beck and Teboulle, 2009]. However, it does have some disadvantages. It is not guaranteed to decrease the objective after each iterate. Further, ISTA may converge rapidly when well-initialized. In our application, this will commonly happen, because parameters are estimated over a range of λ, each solution used as a warm start for the next.

3.5.2   Choosing Tuning Parameters

As of yet we have not discussed how to practically choose the truncation parameters m_1, m_2 and the regularization parameter λ. We suppose the existence of a held-out tuning set; in the absence, one may use cross-validation. We choose m_1, m_2, λ to minimize the negative log-likelihood risk in the held out set. To save on computation, we use the idea of *warm starts* which we detail in the sequel. First, observe that the first-order necessary conditions for the regularized MLE are:

$$\widehat{\mu}_{ij} - \mu_{ij}(\widehat{\theta}_\lambda) - \lambda \widehat{Z}_{ij} = 0, \qquad\qquad i \in V; \qquad\qquad (3.39)$$

$$\widehat{\mu}_i - \mu_i(\widehat{\theta}_\lambda) = 0, \qquad\qquad (i,j) \in E. \qquad\qquad (3.40)$$

where $\widehat{Z}$ denotes the sub gradient of the regularizer $\mathcal{R}$, at $\widehat{\theta}$, which is

$$\widehat{Z}_{ij} = \begin{cases} \{x : \|x\| \leq 1\}, & \|\theta_{ij}\| = 0, \\ \dfrac{\theta_{ij}}{\|\theta_{ij}\|}, & o/w. \end{cases} \tag{3.41}$$

$\widehat{\theta}_{ij} = 0$ when

$$\lambda \geq \|\widehat{\mu}_{ij} - \mu_{ij}(\widehat{\theta}_\lambda)\|. \tag{3.42}$$

When $\widehat{\theta}_{ij} = 0$ for each $(i, j) \in E$, it's clear that $\mu_{ij} = \mu_i \mu_j = \widehat{\mu}_i \widehat{\mu}_j$ by independence and the moment-matching condition (3.40). This allows us to choose an upper bound λ_{max} such that the solution will have no nonzero edge parameters:

$$\lambda_{start} \geq \max_{(i,j) \in E} \|\widehat{\mu}_{ij} - \widehat{\mu}_i \widehat{\mu}_j\|. \tag{3.43}$$

The idea behind warm starting is the following: we begin by estimating $\widehat{\theta}_{\lambda_{start}}$, which amounts to d univariate density estimation problems which can be performed in parallel. Then we fit our model on a path of λ decreasing from λ_{start}, initializing each new problem with the previous solution $\widehat{\theta}_\lambda$. The solution path for the regularized MLE is smooth as a function of λ, suggesting nearby choices of λ will provide values of $\widehat{\theta}$ which are close to one another.

We can also incorporate warm-starting in choosing m_1, m_2. For a given λ, we first estimate the model for first-order polynomials, corresponding to $m_1 = m_2 = 1$. We then increment the truncation parameters by increasing the degree of the polynomial of he sufficient statistics. We augment the previous parameter estimate vector with zeros in the place of the added parameters, and warm start *ISTA* from this vector.

3.6   Experiments

Our simulations were conducted on a workstation operating 23 Intel(R) Xeon(R) CPU E5-2420 1.90GHz processors using `R` with backend computations written in $C++$ and compiled using the `Rcpp` package [Eddelbuettel and François, 2011]. Calculations, including message passing were parallelized using the Threading Building Blocks $C++$ library. We discretized messages uniformly over $[0, 1]$ on a grid of 128 points, and bivariate pseudodensities approximated on a 128×128 grid. All integrals were approximated using Riemann sums over these discretizations. We fit our model using the ISTA algorithm, stopping after an objective improvement of less than 10^{-4} or 1000 iterations. We selected the tuning parameters as described in Section 3.5.2 using warm starts. To choose the edge weights, we generate a series of random spanning trees and take the average edge appearance probability as the edge weight. This produces a valid vector in the spanning tree polytope. We conducted experiments optimizing edge weights by the algorithm in Section 3.4, but found the risk improvement in our simulations to be small relative to the additional computational requirement. We generated three types of data. First we generate independent, identically distributed random Gaussian vectors each with mean $\{0.5, \ldots, 0.5\}$ and covariance Σ. Σ is scaled to have diagonal $1/8^2$ and sparse off-diagonals. The sparsity pattern was generated by randomly including edges with probability $2/d$, so the expected number of edges is $d - 1$. Furthermore, we generate non-Gaussian data by generating Gaussian data by the aforementioned procedure and marginally applying the transformation $y = \text{sign}(x - 0.5) \mid x - 0.5 \mid^{0.6} /5 + 0.5$. That is, the transformed data is distributed as a Gaussian copula. These two models follow a pairwise factorization. Finally, we generated data as a mixture of three non-Gaussian distributions, each with edge structure of a randomly generated spanning tree. The distributions are given equal mixing weights. The resulting mixture of trees is not a pairwise distribution.
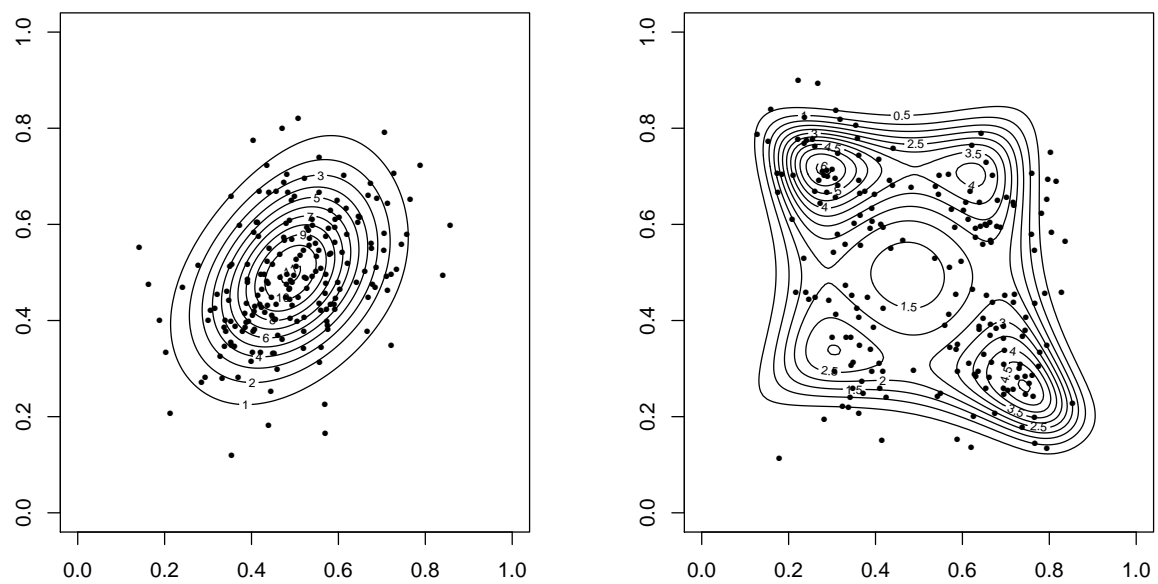
Figure 3.7: Bivariate pseudo density contours estimated from high-dimensional synthetic data. (a) Gaussian data; (b) non-Gaussian (copula) data.

3.6.1   Risk Paths

We first examine the risk paths, varying λ and hence the number of included edges for typical runs of our simulation. We compare the TRW and it's relaxed version (refitting the model under the selected sparsity constraint and setting $\lambda = 0$) to the graphical lasso (using the glasso `R` package) and its relaxed counterpart. Examples of estimated pseudodensities are shown in 3.7. For Gaussian data, we see that TRW performs slightly worse than glasso for sparse graphs, while the gap widens as the graph becomes denser. There are two clear reasons for this. Since the Gaussian model is correct, glasso automatically provides a correctly specified model. TRW must select the number of basis elements and so may include too many (or few) parameters. Furthermore, the variational bound from TRW worsens as the graph becomes denser. For the non-Gaussian data, TRW clearly outperforms the glasso. The risk paths between TRW and glasso look markedly different. TRW demands many more parameters to get the optimal fit, so the performance from the relaxed TRW suffers, while the regularized version benefits by reducing overfitting, thus the relaxed and regular versions have very similar risk. As is well-known, model selection using held-out risk is precarious as the risk path is often flat. In our simulations TRW did quite well. Both glasso and TRW included two false edges for the Gaussian simulation, while TRW omitted several edges. For the non-Gaussian data, glasso included several more false egdes than did TRW.

3.6.2   Density Estimation

We continue by comparing held-out risk estimates to several other high-dimensional density estimation algorithms in Table 3.1: glasso, spherical Gaussian mixture model (with number of components chosen by BIC, [Fraley and Raftery, 2002]), and the Kernel maximum spanning tree estimator of [Liu et al., 2011]. We consider sparse Gaussian, sparse Gaussian copula and tree mixture data for dimensions between 30 to 120. For all simulations we set $n = 100$ and hold out 300 observations for testing. We repeat the simulations (but keeping the generating distribution fixed) 5 times and report the standard errors in parentheses. For

Figure 3.8: Risk path on simulated Gaussian data. Top row is negative log-likelihood risk on held-out data; bottom row are selected graphs. Red edges are false inclusions; gray edges are false omissions.

Gaussian data, TRW performs competitively. It dominates the forest and spherical mixture density estimators but is slightly worse than glasso. For the Gaussian copula data TRW outperforms other methods; TRW can capture both non-Gaussianity and the cyclical dependence structure, while glasso can only capture the latter, and the forest density estimator the former. For mixtures of trees the results are similar to the copula simulation, with TRW dominating the other methods, despite the data not following a pairwise factorization.

	d	Glasso	Glasso Refit	TRW	TRW Refit	Forest	Gaussian Mixture
Gaussian	30	-21.003 (0.175)	-21.22 (0.318)	-20.544 (0.107)	-20.783 (0.244)	-18.969 (0.247)	-19.716 (0.16)
	50	-34.9 (0.185)	-35.136 (0.292)	-34.191 (0.254)	-34.675 (0.258)	-31.242 (0.349)	-32.697 (0.438)
	100	-67.944 (0.19)	-67.698 (0.516)	-67.237 (0.263)	-67.721 (0.334)	-59.91 (0.701)	-65.418 (0.446)
	120	-81.573 (0.531)	-81.441 (0.916)	-80.911 (0.498)	-81.35 (0.649)	-72.763 (1.301)	-78.907 (0.393)
Copula	30	-0.453 (0.191)	-0.508 (0.373)	-3.343 (0.275)	-2.85 (0.236)	-2.128 (0.243)	1.04 (0.16)
	50	-0.1 (0.204)	-0.193 (0.425)	-5.559 (0.279)	-4.817 (0.253)	-3.286 (0.401)	1.802 (0.163)
	100	0.431 (0.295)	0.54 (0.253)	-10.959 (0.393)	-9.91 (0.459)	-5.359 (0.261)	3.396 (0.557)
	120	2.398 (0.243)	2.75 (0.445)	-12.111 (0.434)	-11.283 (0.388)	-4.738 (0.427)	4.313 (0.182)
Mixture of Trees	30	0.406 (0.108)	0.429 (0.142)	-3.215 (0.293)	-2.899 (0.29)	-1.589 (0.27)	1.024 (0.098)
	50	0.67 (0.164)	0.889 (0.306)	-5.217 (0.287)	-4.777 (0.242)	-2.415 (0.444)	1.676 (0.222)
	100	2.5 (0.179)	2.675 (0.251)	-10.461 (0.141)	-9.805 (0.228)	-3.313 (0.365)	3.342 (0.244)
	120	2.837 (0.541)	2.855 (0.499)	-12.365 (0.382)	-11.534 (0.459)	-4.382 (0.408)	4.153 (0.534)

Table 3.1: Held-out risk estimates for synthetic data. Standard deviations for 5 replicates in parenthesis.

3.6.3   ROC Curves

Figures 3.10 and 3.11 display ROC curves for two types of experiments. We generate non-Gaussian data as before, with $n = d = 30$. The curves trace the true negative and true positive percentages of the algorithms, varying the regularization parameter λ. We choose m_1, m_2 so that the resulting curve has the largest value of $\max_\lambda TP(\lambda) + TN(\lambda)$. The plotted curves are an average of 20 repetitions. In Figure 3.10 the edges are chosen to be a randomly-generated spanning tree; in Figure 3.11 the edges are included with equal probability $2/d$, which we call the ER graph. We compare the TRW estimator to the SKEPTIC estimator from [Liu et al., 2012a]. The SKEPTIC estimator was particularly devised for estimating Gaussian copula graphical models, while our estimator is designed for a superset of those models.

Overall the SKEPTIC performs better in terms of area under the curve (AUC). It also generally has a better TP% for moderate or small values of TN%. However the TRW estimator still performs quite well in these two metrics. For large TN% the TRW estimator manages a better TP%. We can only speculate on why the TRW estimate performs better here, but it may be because when the selected graph is very sparse it has few or no cycles, and the tree-based approximation of TRW is more powerful. This is consistent with the experiments, as the phenomenon is more pronounced for the tree simulation than the loopy graph simulation.

3.6.4   MEG Data

Magnetoencephalography or MEG is a neuroimaging technique for mapping brain activity using electrical currents in the brain. The resulting signals are high-frequency and have a complex non-linear relation to one another. There has been interest in using various neuroimaging techniques for mapping regional brain networks [Kramer et al., 2011], and particularly in understanding differences in connectivity related to neurodegenerative diseases [Stam, 2010]. We explore this on the MEG data from [Vigário et al., 1998], which
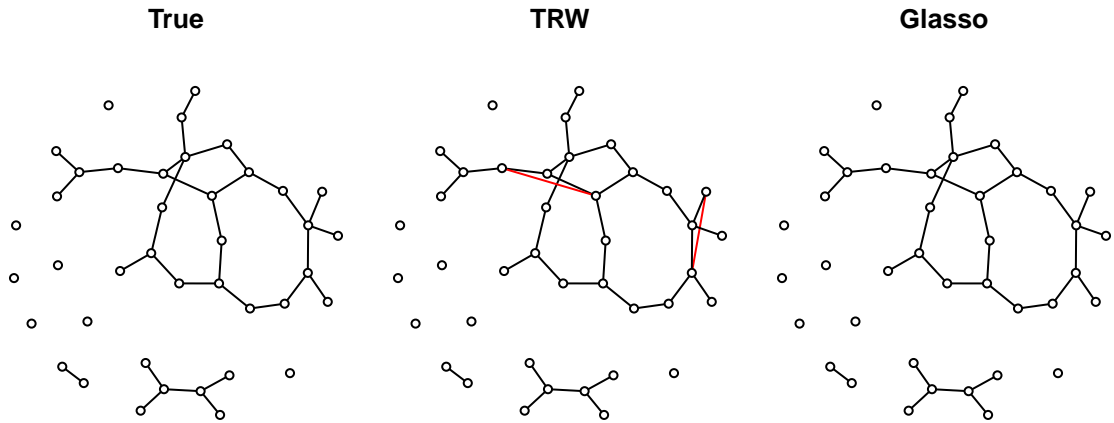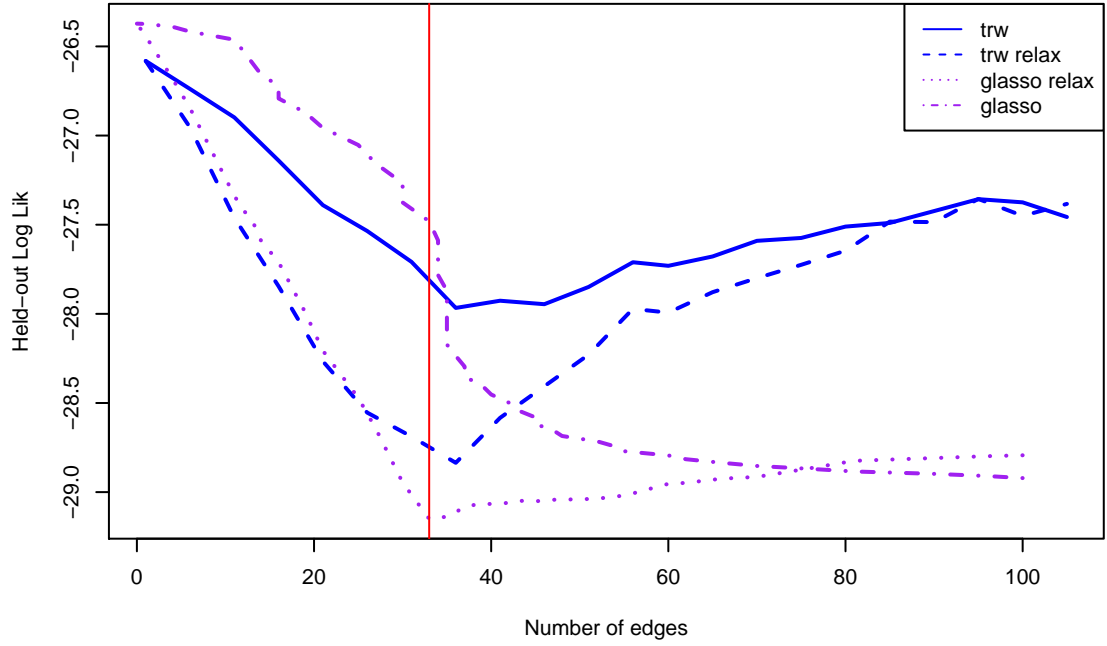
47
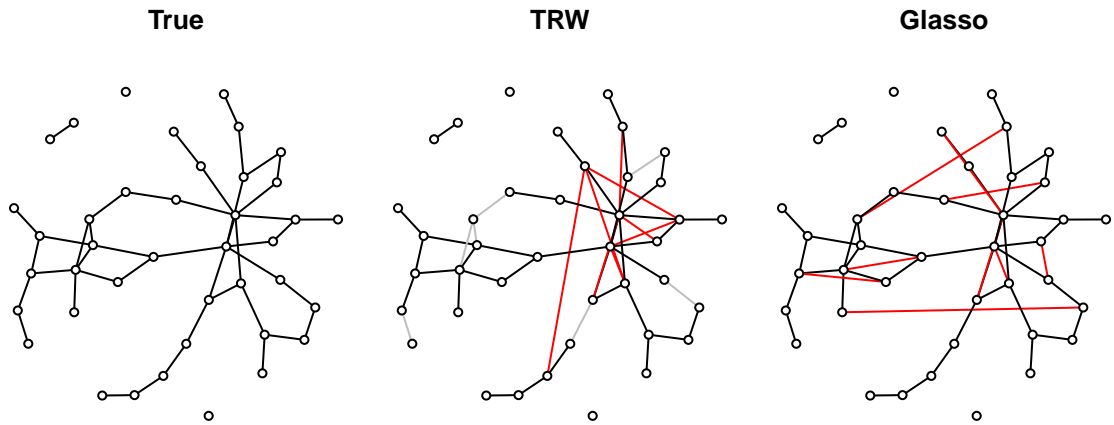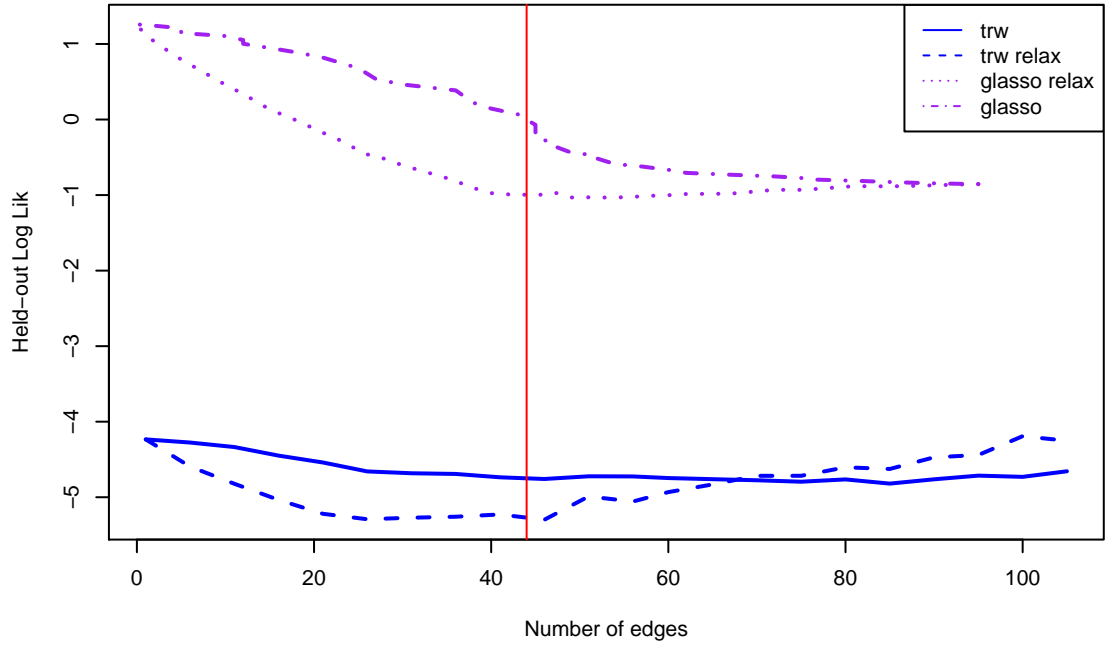
Figure 3.9: Performance on simulated Gaussian copula data. Top row is negative log-likelihood risk on held-out data; bottom row are selected graphs. Red edges are false inclusions; gray edges are false omissions.
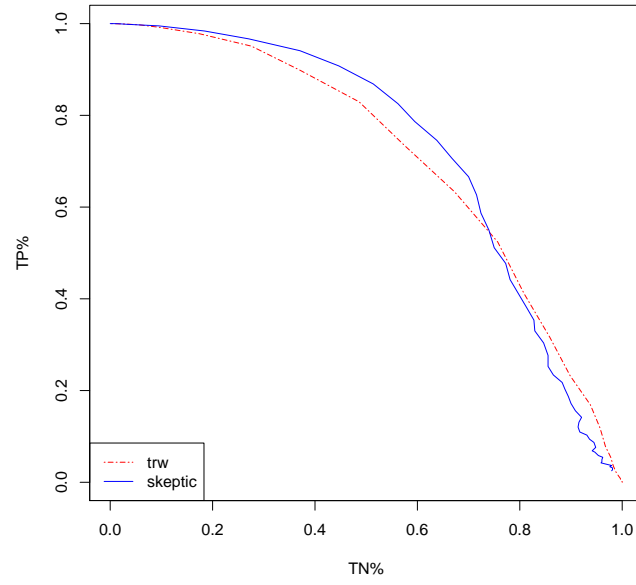
Figure 3.10: ROC Curve, tree graph, n=d=30. Solid line: SKEPTIC estimator. Dotted line: TRW.
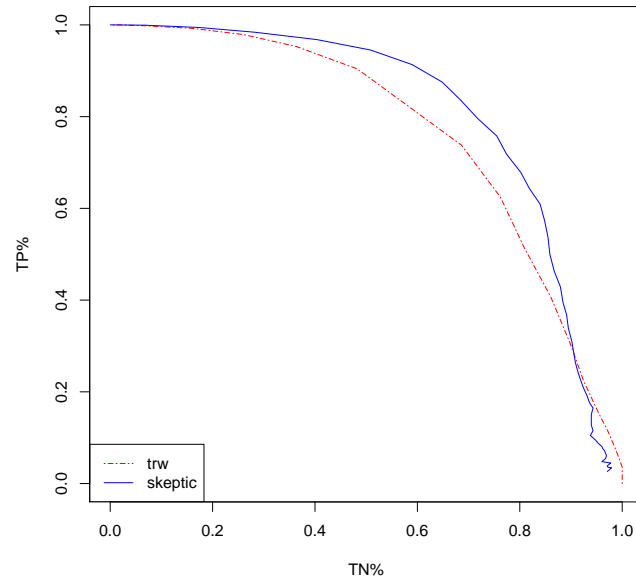
Figure 3.11: ROC Curve, ER graph, n=d=30. Solid line: SKEPTIC estimator. Dotted line: TRW.

contains measurements from 122 sensors. We scale the data to be contained in the unit cube and remove large outliers (marginally larger than 6 standard deviations). Two features of this data are the temporal dependence of the signals and the presence of artifacts. Since our main motivation is graph estimation we will not address these issues, besides restricting our attention to a small timespan of the dataset. We use the first 400 observations in the series, randomly assigning 100 as training data, and the other 300 as test data.

Inspecting two-dimensional projections of the data and their corresponding pseudodensities in Figure 3.12, the data displays clearly non-linear and multi-modal behavior which TRW can capture, but glasso cannot. We compare the glasso and TRW methods for graph estimation. We select a graph to minimize the held-out risk for the relaxed TRW, which has 199 edges. We then estimate the graph using the graphical lasso, setting the regularizaition parameter to include the same number of edges. The estimated graphs are shown in figure 3.13. For clarity, we color code vertices from the top four clusters produced from running the learned graphs through the community detection algorithm of [Newman, 2006]. Note that the position of vertices here does not correspond to the actual location of the sensors on the scalp. Comparing the graphs from the minimum risk estimators, the two graphs share many features in common, each having one large connected component containing several smaller densely connected communities. However, they have clear differences, disagreeing on 66 edges, or one-third of the edges in each respective graph. We believe due to the complex nature of the observed signals in imaging data such as MEG, our method may be able to bring new and better insights to understanding functional connectivity of brain networks.

3.7   Discussion

In this section we detail a tree-reweighted variational approximation for continuous-valued exponential families, which we apply to the exponential series estimator of chapter 1. Evaluating the variational likelihood involves message passing which can be effectively parallelized for high-dimensional problems, and provides a lower-bound to the likelihood (and upper
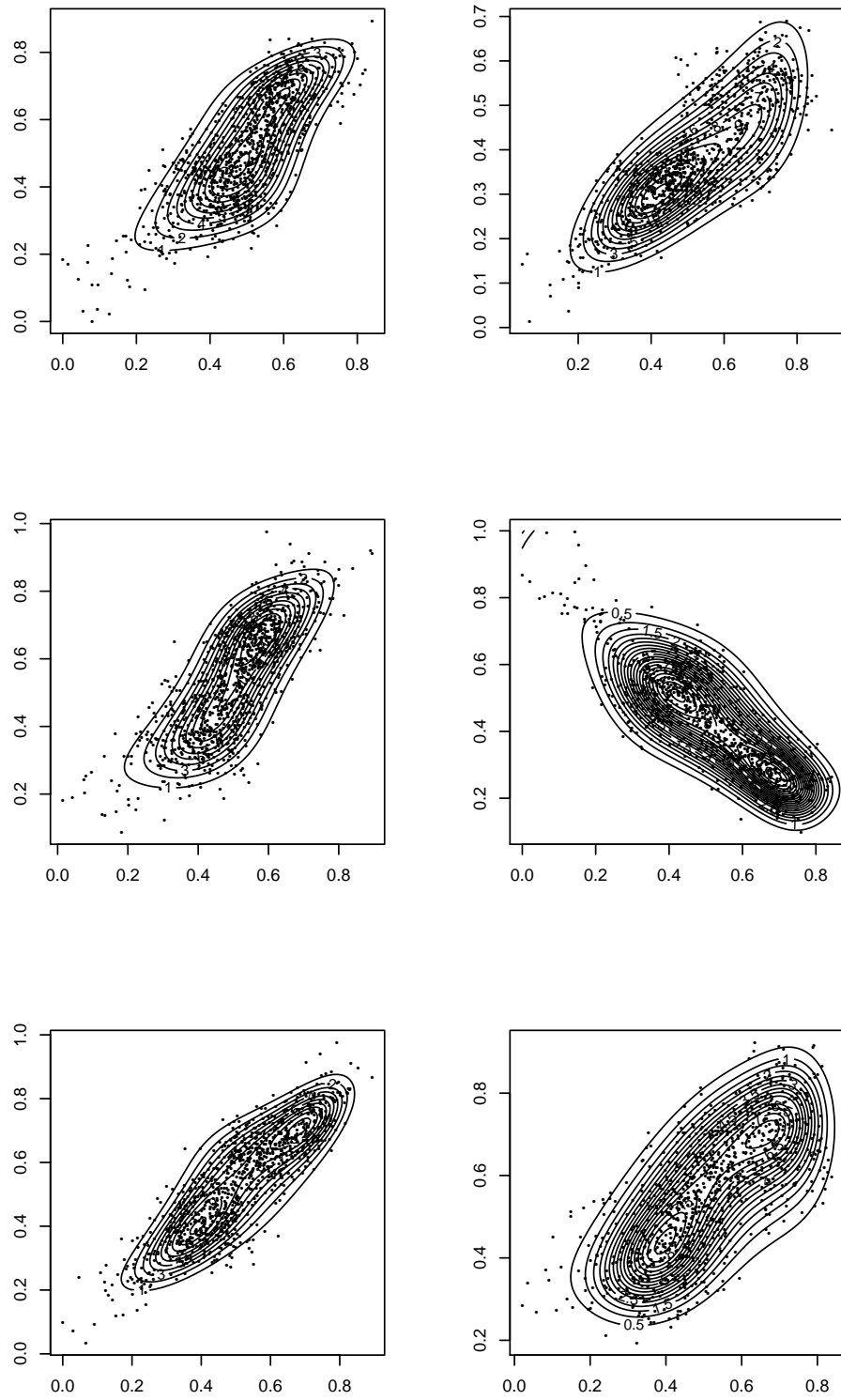
Figure 3.12: Six estimated pseudodensities plotted with 1000 observations of the MEG data.
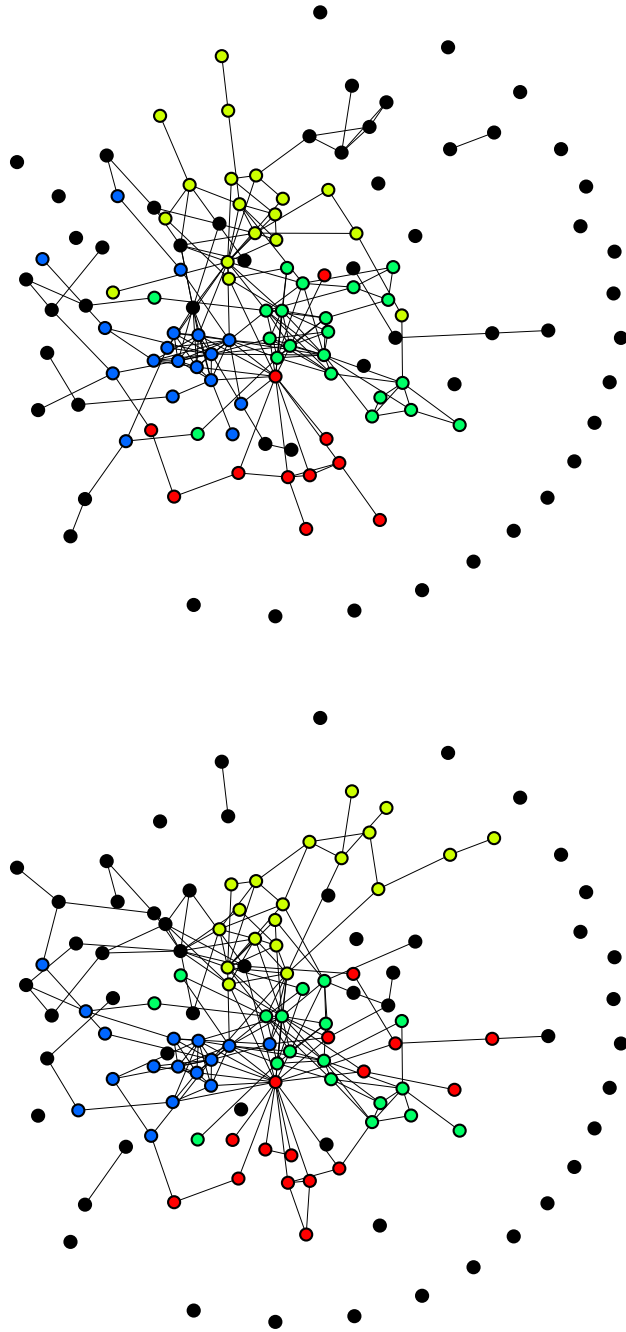
Figure 3.13: Estimated graphs from MEG data. Top: Graph estimated from TRW ; Bottom: Graph learned from glasso

bound to risk). We describe a proximal gradient algorithm for estimating the regularized MLE. Our experiments show this approach has very attractive performance in both risk and model selection performance, compared to other methods in the literature. We also demonstrate our method on a data set of MEG signals.

CHAPTER 4

REGULARIZED SCORE MATCHING

4.1 Introduction

Undirected graphical models are an invaluable class of statistical models. They have been used successfully in fields as diverse as biology, natural language processing, statistical physics and spatial statistics. The key advantage of undirected graphical models is that its joint density may be factored according to the cliques of a graph corresponding to the conditional dependencies of the underlying variables. The go-to approach for statistical estimation is the method of maximum likelihood (MLE). Unfortunately, with few exceptions, MLE is intractable for high-dimensional graphical models, as it requires computation of the normalizing constant of the joint density, which is a d-fold convolution. Even exponential family graphical models [Wainwright and Jordan, 2008], which are the most popular class of parametric models, are generally non-normalizable, with a notable exception being the Gaussian graphical model. Thus the MLE must be approximated. State of-the-art methods for graphical structure learning avoid this problem by performing neighborhood selection [Yang et al., 2012; Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010]. However, this approach only works for special types of pairwise graphical models whose conditional distributions form GLMs. Furthermore, these procedures do not by themselves produce parameter estimates.

In this chapter we demonstrate a powerful new method for graph structure learning and parameter estimation based on minimizing the regularized Hyvärinen score of the data. It works for any continuous pairwise exponential family, as long as it follows some weak smoothness and tail conditions. Our method allows for multiple parameters per vertex/edge. We prove high-dimensional model selection and parameter consistency results, which adapt to the underlying sparsity of the natural parameters. As a special case, we derive a new method for estimating sparse precision matrices with very competitive estimation and graph

learning performance. We also consider how our method can be used to do model selection for the general nonparametric pairwise model by choosing the sufficient statistics to be basis elements with degree growing with the sample size. We show our method can be expressed as a second-order cone program, and which we provide highly scalable algorithms based on ADMM and coordinate-wise descent.

4.2   Background

4.2.1   Graphical Models

Suppose $X = (X_1, \ldots, X_d)$ is a random vector with each entry having support $\mathcal{X}_i$, $i = 1, \ldots, d$. Let $G = (V, E)$ be an undirected graph on d vertices corresponding to the elements of X. An undirected graphical model or Markov random field is the set of distributions which satisfy the Markov property or condition independence with respect to G. From the Hammersley-Clifford theorem, if X is Markov with respect to G, the density of X, p can be decomposed as

$$p(x) \propto \exp \left\{ \sum_{c \in \mathrm{cl}(G)} \psi_c(x_c) \right\}, \tag{4.1}$$

where $\mathrm{cl}(G)$ is the collection of cliques of G. The pairwise graphical model supposes the density can be further factored according to the edges of G,

$$p(x) \propto \exp \left\{ \sum_{i,j \in V, i \leq j} \psi_{ij}(x_i, x_j) \right\}. \tag{4.2}$$

For a pairwise exponential family, we parametrize ψ_i, ψ_{ij} by

$$\psi_{ii}(x_i) := \sum_{u \leq m} \theta_i^u \phi_i^u(x_i), \qquad\qquad i \in V, \qquad (4.3)$$

$$\psi_{ij}(x_i, x_j) := \sum_{u \leq m} \theta_{ij}^u \phi_{ij}^u(x_i, x_j), \qquad\qquad (i,j) \in E. \qquad (4.4)$$

Here m denotes the maximum number of statistics per edge or vertex. We denote θ to be the vectorization of the parameters, $\theta := (\theta_{11}^\top, \ldots, \theta_{d1}^\top, \theta_{22}^\top, \ldots, \theta_{dd}^\top)^\top$.

4.2.2 Scoring Rules

A *scoring rule* [Dawid and Lauritzen, 2005] $S(x, Q)$ is a function which measures the predictive accuracy of a distribution Q on an observation x. A scoring rule is *proper* if $\mathbb{E}_p[S(X,Q)]$ is uniquely minimized at $Q = P$. When Q has a density q, we equivalently denote the scoring rule $S(X, q)$. A *local* scoring rule only depends on q through its evaluation at the observation x. A proper scoring rule induces an entropy

$$H(p) = \mathbb{E}_p\left[S(X, p)\right], \qquad (4.5)$$

as well as a divergence

$$D(p, q) = \mathbb{E}_p\left[S(X, q) - S(X, p)\right]. \qquad (4.6)$$

An optimal score estimator is an estimator which minimizes the empirical score

$$\frac{1}{n} \sum_{r=1}^{n} S(X^r, q), \qquad (4.7)$$

over some class of densities.

Example 4.2.1. The *log score* takes the form $l(x, q) := -\log q(x)$. The corresponding

entropy is the Shannon entropy $H(p) := -\mathbb{E}_p [\log p]$, its corresponding divergence is the *Kullback-Leibler Divergence* $\mathrm{KL}(p \mid q) = \mathbb{E}_{X \sim p} \left[\log \frac{p(X)}{q(X)} \right]$ and the optimal score estimator is the maximum likelihood estimator. It is a proper and local scoring rule. This scoring rule was implemented in Chapter 2.

Example 4.2.2. Consider the *Bregman score*,

$$b(x, q) := -g'(q(x)) - \int \rho(dy) \left(g(q(y)) - q(y) g'(q(y)) \right), \tag{4.8}$$

where $g : \mathbb{R}^+ \to \mathbb{R}$ is a convex, differentiable function and ρ is some baseline measure. The corresponding entropy is $H(p) = - \int \rho(dy) g(p)$ and divergence

$$D(p, q) = \int \rho(dy) \left(g(p) - \left(g(q) + g'(q)(p - q) \right) \right). \tag{4.9}$$

When $g(x) = \log(x)$, after removing the constant term, b has the form

$$b(x, q) = -\frac{1}{q(x)} - \int \rho(dy) \log(q(y)) \tag{4.10}$$

$$= -e^{-f(x)} - \int \rho(dy) f(y), \tag{4.11}$$

where $f = \log q$. This is a proper scoring rule [Dawid and Musio, 2014], but it is not local because it depends on values of q besides the observation x. An estimation procedure for nonparametric graphical models using smoothing splines was based on this scoring rule in [Jeon and Lin, 2006].

4.2.3 Hyvärinen Score

Consider densities q which are twice continuously differentiable over $\mathcal{X} = \mathbb{R}^d$ and satisfy

$$\|p(x) \nabla \log q(x)\| \to 0, \text{ for all } \|x\| \to \infty. \tag{4.12}$$

where $X \sim p$. Consider the scoring rule

$$h(x, q) = \frac{1}{2}\|\nabla \log q(x)\|_2^2 + \Delta \log q(x), \qquad (4.13)$$

where ∇ denotes the gradient operator and Δ is the operator

$$\Delta \phi(x) = \sum_{i \in V} \frac{\partial^2 \phi(x)}{\partial x_i^2}. \qquad (4.14)$$

This is a proper and local scoring rule [Parry et al., 2012]. Using integration by parts, it can be shown it induces the *Fisher divergence*:

$$\mathrm{F}(p \mid q) = \mathbb{E}_{X \sim p}\left[\left\|\nabla \log \frac{p(X)}{q(X)}\right\|_2^2\right]. \qquad (4.15)$$

The optimal score estimator is called the *score matching* estimator [Hyvärinen, 2005, 2007]. The Hyvärinen score is homogeneous in q [Parry et al., 2012], so that it does not depend on the normalizing constant of q, which for multivariate exponential families is typically intractable. Second, for natural exponential families the objective of the optimal score estimator is quadratic, so the estimating equations corresponding to score matching are linear in the natural parameters [Forbes and Lauritzen, 2014]. Maximum likelihood for exponential families generally involves a complex mapping from the sufficient statistics of the data to the natural parameters [Wainwright and Jordan, 2008; Brown, 1986], necessitating specialized solvers.

4.2.4   Score Matching for Exponential Families

For a pairwise density, define $\phi_{.,i} = \left(\phi_{1i}^\top, \ldots, \phi_{di}^\top\right)^\top$. For $i \in V$, denote $a_i(x) := \frac{\partial}{\partial x_i}\phi_{.,i}$ and

$$(K(x))_{.,i} := \frac{\partial^2 \phi_{.,i}}{\partial x_i^2}. \qquad (4.16)$$

Taking derivatives,

$$\frac{\partial}{\partial x_i} \langle \phi(x), \theta \rangle = \left\langle \frac{\partial \phi_{\cdot,i}}{\partial x_i}, \theta_{\cdot,i} \right\rangle, \tag{4.17}$$

thus h takes the form

$$h(x, \theta) = \sum_{i \in V} \left(\frac{1}{2} \theta_{\cdot,i}^\top a_i(x) a_i(x)^\top \theta_{\cdot,i} + K_{\cdot,i}(x)^\top \theta_{\cdot,i} \right). \tag{4.18}$$

h is a sum of d positive-semidefinite quadratic forms, so it is also psd quadratic. Alternatively, we may write $h(x, \theta) = \theta^\top A(x)\theta + K(x)^\top \theta$, where $A(x)$ is a psd matrix with at most $2md$ non-zero entries per row, and $K(x)$ is a vector with $K_{ij} = \frac{\partial^2 \phi_{ij}}{\partial x_i^2} + 1\{i \neq j\} \frac{\partial^2 \phi_{ij}}{\partial x_j^2}$. If we write $\tilde\theta = \left(\theta_{\cdot,1}^\top, \theta_{\cdot,2}^\top, \ldots, \theta_{\cdot,d}^\top\right)^\top$, where $\tilde\theta_{ij} = \tilde\theta_{ji}$, we may write the scoring rule as

$$h(x, \tilde\theta) = \frac{1}{2} \tilde\theta^\top \tilde A(x) \tilde\theta + \tilde K(x)^\top \tilde\theta, \tag{4.19}$$

where $\tilde A(x)$ is a block-diagonal matrix,

$$\tilde A(x) = \begin{pmatrix} a_1(x)a_1(x)^\top & & & \\ & a_2(x)a_2(x)^\top & & \\ & & \ddots & \\ & & & a_d(x)a_d(x)^\top \end{pmatrix} \tag{4.20}$$

and $\tilde K = (K_{\cdot,1}^\top, \ldots, K_{\cdot,d}^\top)^\top$. We will alternate between these two equivalent representations of h based on convenience.

Remark 4.2.1 (Bounded supports). From the differentiability assumption we see that our derivations do not generally apply when $\mathcal{X}_i$ is a half-bounded or bounded support as the density may not be differentiable at the boundary. However, in [Hyvärinen, 2007] a proper scoring rule was derived for half-bounded supports, which may be shown to have the same

form as (4.19), after modifying slightly the formulas for $a_i(x), K(x)$. Here we derive a similar formula for densities on $[0,1]^d$.

Proposition 4.2.2. Consider random vectors taking values in $[0,1]^d$, with density q; suppose $X \sim p$. If q is twice continuously differentiable and satisfies

$$\|p(x)\nabla \log q(x) \otimes x(1-x)\| \to 0, \text{ for all } x \text{ approaching the boundary,} \qquad (4.21)$$

where $\otimes$ denotes the tensor product $x \otimes y := (x_1 y_1, \ldots, x_d y_d)$, then

$$
\begin{aligned}
h(x,q) &:= \frac{1}{2}\|\nabla \log q(x) \otimes x(1-x)\|_2^2 \\
&\quad + \sum_{i \in V}\left(-2(2x_i - 1)x_i(1-x_i)\frac{\partial \log q(x)}{\partial x_i} + x_i(1-x_i)\frac{\partial^2 \log q(x)}{\partial x_i^2}\right), \qquad (4.22)
\end{aligned}
$$

is a proper scoring rule. In particular when q is an exponential family with natural parameters θ and sufficient statistics ϕ, $h(x,\tilde\theta) = \frac{1}{2}\tilde\theta^\top \tilde A(x)\tilde\theta + K(x)^\top \theta$ is a proper scoring rule, where

$$K(x)_{ij} = -2(2x_i - 1)x_i(1-x_i)\frac{\partial \phi_{ij}}{\partial x_i} + (x_i(1-x_i))^2\frac{\partial^2 \phi_{ij}}{\partial x_i^2}, \qquad (4.23)$$

$$a_i(x) = x_i(1-x_i)\frac{\partial \phi_{\cdot,i}}{\partial x_i}. \qquad (4.24)$$

and $\tilde A(x) = diag(a_i(x)a_i(x)^\top)$.

Thus, all of the results in this work may be effortlessly carried over to exponential families over bounded supports.

4.3   Previous Work

There is a small but growing literature on applications using the Hyvärinen score for estimation. [Sriperumbudur et al., 2013] consider using the Hyvärinen score for density estimation in a reproducing kernel Hilbert space (RKHS). They consider the optimization for a density

q,

$$\min_{q} \left\{ \frac{1}{n} \sum_{k=1}^{n} h(X^k, q) + \frac{\lambda}{2} \|q\|_{\mathcal{H}}^2 \right\}, \tag{4.25}$$

where $\|\cdot\|_{\mathcal{H}}^2$ is the norm of the RKHS. After an application of the representer theorem [Kimeldorf and Wahba, 1971], they show this may be expressed as a finite-dimensional quadratic program. They derive rates for convergence to the true density with respect to the Fisher divergence.

[Vincent, 2011] shows that the denoising autoencoder may be expressed as a type of score matching estimator, which they call *denoising score matching*. Suppose that $\tilde{X}$ is a version of a sample X which has been corrupted by Gaussian noise, so that its conditional distribution has the score $\partial \log q(\tilde{x} \mid x) = \frac{1}{\sigma^2}(x - \tilde{x})$. Suppose we seek to fit the corrupted data according to a density of the form

$$\log p(\tilde{x} \mid W, b, c) \propto -\frac{1}{\sigma^2} \left(\langle c, \tilde{x} \rangle - \frac{1}{2} \|\tilde{x}\|_2^2 + \text{softplus} \left(\sum_j \langle W_j, \tilde{x} \rangle + b_j \right) \right), \tag{4.26}$$

where $\text{softplus}(x) = \max(0, x)$, then minimizing the Fisher divergence between the model density and $q(\tilde{x} \mid x)$ can be shown to be equivalent to minimizing

$$\mathbb{E}_{q(\tilde{x}, x)} \left[\|W^\top \text{sigmoid}(W\tilde{X} + b) + c - X\|^2 \right]. \tag{4.27}$$

This is a simple denoising autoencoder with a single hidden layer, encoder $f(\tilde{x}) = \text{sigmoid}(W\tilde{x} + b)$, and decoder $f'(y) = W^\top y + c$.

Score matching has also been used for learning natural image statistics [Kingma and LeCun, 2010; Köster et al., 2009].

4.4   Score Matching Estimator

Define the statistics

$$\widehat{\Gamma} = \frac{1}{n} \sum_{r=1}^{n} A(X^r), \tag{4.28}$$

$$\widehat{K} = \frac{1}{n} \sum_{r=1}^{n} K(X^r). \tag{4.29}$$

The *regularized score matching estimator* is a solution to the problem

$$\widehat{\theta} \in \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{2} \theta^{\top} \widehat{\Gamma} \theta + \widehat{K}^{\top} \theta + \mathcal{R}(\theta) \right\}. \tag{4.30}$$

Here $\mathcal{R}$ is the group penalty

$$\mathcal{R}(\theta) = \sum_{i,j \in V} \|\theta_{ij}\|_2. \tag{4.31}$$

This norm induces sparsity in groups (i.e. edges/vertices). In high dimensions, regularizing the vertex parameters is necessary, as (4.30) need not exist otherwise. Both the scoring rule and regularizer of (4.30) are convex in θ, so it is a convex program. In particular, observe that it can be equivalently represented as

$$\min_{t, t_{ij}} \left\{ t + \lambda \sum_{ij} t_{ij} \right\} \tag{4.32}$$

$$s.t. \quad t \geq \frac{1}{2} \theta^{\top} \widehat{\Gamma} \theta + \widehat{K}^{\top} \theta,$$

$$t_{ij} \geq \|\theta_{ij}\|.$$

(4.32) is a second-order cone program (SOCP) [Boyd and Vandenberghe, 2004], as the quadratic constraint can be re-written as a conic constraint. If $\widehat{\Gamma}$ is not positive definite, particularly when $n > d$, (4.30) may not be unique. This is typical for high dimensional

problems. One can impose further assumptions to guarantee uniqueness. For example various assumptions have been described for the lasso (see an overview of these assumptions in [Tibshirani et al., 2013]) , but we won't go into those details here.

4.4.1 Gaussian Score Matching

Consider the Gaussian density:

$$q(x) \propto \exp\left\{-\frac{1}{2}x^\top \Omega x\right\}, \tag{4.33}$$

for $\Omega \succ 0$. We have

$$\nabla \log q(x) = -\Omega x, \tag{4.34}$$

$$\nabla_i(\nabla_i \log q(x)) = -\Omega_{ii}. \tag{4.35}$$

so the Hyvärinen score is given by

$$h(x, \Omega) = -\sum_i \Omega_{ii} + \frac{1}{2}x^\top \Omega^\top \Omega x \tag{4.36}$$

$$= \operatorname{trace}\left(-\Omega + \frac{1}{2}\Omega^2 xx^\top\right). \tag{4.37}$$

Let $\widehat{\Sigma} = \frac{1}{n}\sum_{r=1}^n X^r(X^r)^\top$. The optimal regularized score estimator $\widehat{\Omega}$ is the solution to

$$\min_{\Omega=\Omega^\top}\left\{\operatorname{trace}\left(\frac{1}{2}\Omega\widehat{\Sigma}\Omega - \Omega\right) + \lambda\|\Omega\|_1\right\}. \tag{4.38}$$

In the notation of (4.30), we have $\theta = \operatorname{vec}(\Omega)$, $\widehat{K} = \operatorname{vec}(I_d)$ and $\widehat{\Gamma}_i = \widehat{\Sigma}$ for each $i \in V$. We do not impose a positive definite constraint on Ω. Doing so would still result in a convex program, indeed it is a semidefinite program, but the resulting computation becomes more complicated and less scalable in practice. However, our theoretical results imply that $\widehat{\Omega}$ is positive definite with high probability. Indeed, denote $\|\widehat{\Omega} - \Omega^*\|_{sp}$ the spectral norm

(maximum absolute value of eigenvalues) of the difference $\widehat{\Omega} - \Omega^*$. Since the spectral norm is dominated by the Frobenius norm (elementwise L_2 norm), the consistency result in the sequel implies consistency in spectral norm, and so the eigenvalues of $\widehat{\Omega}$ will be positive with probability approaching one, assuming the population precision matrix Ω^* has strictly positive eigenvalues. Furthermore, we note that our model selection guarantees still follow whether or not the estimator $\widehat{\Omega}$ is positive definite.

4.5   Main Results

We suppose we are given i.i.d. data $X^1, \ldots, X^n \sim p^*$. p^* need not belong to the pairwise exponential family being estimated, in which case we may think of our consistency results as being relative to the population quantity

$$\theta^* := \left\{ \mathbb{E}_{p^*}[A(X)] \right\}^{-1} \mathbb{E}_{p^*}[K(X)] \tag{4.39}$$

$$= (\Gamma^*)^{-1} K^*. \tag{4.40}$$

Define the maximum column sum of θ^* by

$$\kappa_{\theta,1} := \max_{i \in V} \sum_{j \in V, u \leq m} (\theta^*)_{ij}^u, \tag{4.41}$$

and define the maximum degree as

$$s := \max_{i \in V} |\{(i,j) : (i,j) \in E\}|. \tag{4.42}$$

Assumption 4.5.1. $\Gamma_i^* = \mathbb{E}_{p^*}[a_i(X)a_i(X)^\top]$ satisfies for each $i \in V$,

$$\infty > \bar{\epsilon} \geq \Lambda_{\max}(\Gamma_i^*) \geq \Lambda_{\min}(\Gamma_i^*) \geq \underline{\epsilon} > 0. \tag{4.43}$$

Note that this also implies that the eigenvalues of Γ^* are bounded as the rows of Γ^* are

non-trivial linear combinations of those of $\text{diag}(\Gamma_i^*)$, so the inverse in (4.40) exists and is unique.

We also suppose θ^* is sparse, in the following sense:

Assumption 4.5.2. θ^* belongs to the set

$$\tilde{\mathcal{P}} := \tilde{\mathcal{P}}(E) = \left\{ \theta : \|\theta_{ij}\|_2 = 0, \text{ for } (i,j) \in E^c \right\}. \tag{4.44}$$

For both parameter consistency and model selection we require the following tail conditions:

Assumption 4.5.3. For each $i, j, k \in V$ and $u \leq m$ and $t \leq \nu$, for some $c_1, c_2, \nu > 0$,

$$\mathbb{P}\left(\left| \widehat{K}_{ij}^u - (K^*)_{ij}^u \right| \geq t \right) \leq \exp\left\{ -c_1 n t^2 \right\} \tag{4.45}$$

$$\mathbb{P}\left(\left| (\widehat{\Gamma}_i)_{jk}^u - (\Gamma_i^*)_{jk}^u \right| \geq t \right) \leq \exp\left\{ -c_2 n t^2 \right\}. \tag{4.46}$$

4.5.1 Parameter Consistency

We present results in terms of the (vector) L_2 norm. Note in particular that this result doesn't require any incoherence condition (though we do require for model selection consistency in the sequel).

For the parameter consistency results in particular, we require the following sub-Gaussian assumption:

Assumption 4.5.4. For each $i \in V$ and $r = 1, \ldots, n$, $a_i(X^r)$ is a sub-Gaussian random vector.

Theorem 4.5.5. *Suppose the regularization parameter is chosen as*

$$\lambda_n \asymp \sqrt{\frac{m \kappa_{1,\theta}^2 \log(md)}{n}}, \tag{4.47}$$

65

if the sample size satisfies

$$n = \Omega(md), \tag{4.48}$$

then any solution to regularized score matching satisfies

$$\|\widehat{\theta} - \theta^*\|_2 = O_p \left(\sqrt{\frac{(d + |E|)m\kappa_{1,\theta}^2 \log(md)}{n}} \right). \tag{4.49}$$

Remark 4.5.6. Consider Gaussian score matching. Here $m = 1$, so if $\kappa_{1,\theta}$ is bounded we have $\|\widehat{\theta} - \theta^*\|_2 = O_p \left(\sqrt{\frac{(d+|E|)\log(d)}{n}} \right)$. This rate is the same as the graphical lasso shown in [Rothman et al., 2008]. Furthermore, here $\Gamma_i^* = \Sigma$, so our assumption 4.5.1 amounts to bounds on the eigenvalues of Σ, which are the same as for sparse precision matrix MLE. The assumption that $\kappa_{1,\theta}$ is bounded here says that the sums of the absolute value of rows of Ω^* are bounded, which is not necessary for the regularized MLE.

Remark 4.5.7. We might reasonably expect $\kappa_{1,\theta} = O(sm)$, in which case $\|\widehat{\theta} - \theta^*\|_2 = O_p \left(\sqrt{\frac{(d+|E|)m^3 s^2 \log(md)}{n}} \right)$. In this setting the regularized MLE will have the rate

$$O_p \left(\sqrt{\frac{(d + |E|)m \log(md)}{n}} \right). \tag{4.50}$$

(see results in Appendix A).

4.5.2 Model Selection

For model selection we require several additional conditions. Denote $\widehat{E}$ as the edge set learned from $\widehat{\theta}$:

$$\widehat{E} := \left\{ (i, j) : \|\widehat{\theta}_{ij}\|_2 = 0 \right\}. \tag{4.51}$$

Furthermore, define

$$\kappa_\Gamma := \|(\Gamma^*)^{-1}\|_\infty, \tag{4.52}$$

$$\kappa_\theta := \|\theta^*\|_{\max}, \tag{4.53}$$

$$\rho^* := \min_{(i,j)\in E} \|\theta_{ij}\|_{\max}. \tag{4.54}$$

Here $\|A\|_\infty = \max_j \sum_i |A_{ij}|$ is the matrix ∞ norm and $\|\cdot\|_{\max}$ the elementwise max norm. We require an *incoherence condition*:

Assumption 4.5.8.

$$\max_{(i,j)\in E^c} \|\Gamma^*_{ij,E}(\Gamma^*_{EE})^{-1}\|_2 \leq \frac{1-\tau}{\sqrt{d+E}}, \qquad \text{for some } \tau \in (0,1]. \tag{4.55}$$

where $\|A\|_2$ is the matrix operator norm.

In the following theorem we suppose $\kappa_\Gamma, \kappa_\theta, s$ are are bounded, while ρ^* may change with the sample size.

Theorem 4.5.9. *Suppose the regularization parameter λ_n is chosen to be*

$$\lambda_n \asymp \sqrt{\frac{m\kappa^2_{1,\theta}\log(dm)}{n}}, \tag{4.56}$$

then if

$$n = \Omega(\max\{m\kappa^2_{1,\theta}\log(dm), m^2 s^2 \log(dm)\}), \tag{4.57}$$

$$\frac{1}{\rho^*} = o\left(\sqrt{\frac{\kappa^2_{1,\theta}\log(dm)}{n}}\right), \tag{4.58}$$

there exists a solution to the regularized score matching estimator $\widehat{\theta}$ with estimated edge set

$\widehat{E}$ *satisfying*

$$\mathbb{P}(\widehat{E} = E) \to 1. \tag{4.59}$$

Remark 4.5.10. Assuming $m, s, \kappa_{1,\theta}$ are bounded, this implies the dimension may grow nearly exponentially with the sample size:

$$d = o(e^n), \tag{4.60}$$

with the probability of model selection consistency still aproaching one.

Remark 4.5.11 (Gaussian score matching). When $m = 1$, the sample complexity matches that for structure learning of the precision matrix using the log-det divergence, in [Ravikumar et al., 2011]. Thus Gaussian score matching in particular benefits from identical model selection guarantees as the graphical lasso algorithm. However it should be noted that the assumptions are slightly different. In particular the graphical lasso requires an irrepresentable condition on $\Sigma \otimes \Sigma$, while our method involves an irrepresentable condition for $\Sigma \otimes I_d$.

4.5.3 Model Selection for the Nonparametric Pairwise Model

In this section we consider model selection for the nonparametric pairwise model. We suppose the log of the true density p^* belongs to W_2^r, the Sobolev space of order r. This implies, along with the pairwise assumption, that $\log p^*$ has the infinite expansion

$$\log p^* \propto \exp \left\{ \sum_{i,j \in V, i \leq j} \sum_{k,l=1}^{\infty} (\theta^*)_{ij}^{kl} \phi_{kl}(x_i, x_j) + \sum_{i \in V} \sum_{k=1}^{\infty} (\theta^*)_i^k \phi_k(x_i) \right\}, \tag{4.61}$$

where here $\{\phi_k, \phi_{kl}\}$ is a basis over $[0,1]^2$. For an expansion in W_2^r, we have that the coefficients decay at the following rates:

$$\sum_k \theta_i^k k^{2r} < \infty, \qquad\qquad \text{for all } i \in V, \qquad\qquad (4.62)$$

$$\sum_{k,l} \theta_{ij}^{kl} k^{2r_i} l^{2r_j} < \infty, \qquad\qquad \text{for all } (i,j) \in E, \qquad r_i + r_j = r. \qquad (4.63)$$

For our results we assume $\{\phi_k\}$ is the orthonormal Legendre basis on $[0,1]$, and $\{\phi_{kl}\}$ is the tensor product basis $\phi_{kl}(x_i, x_j) = \phi_k(x_i) \cdot \phi_l(x_j)$. This is because the supporting lemmas are particular to the Legendre basis, but in practice one is not limited to a particular basis. Now, consider forming a density by truncating (4.61) after m_1 terms for the univariate expansions, and m_2 for bivariate:

$$\log p_\theta \propto \exp\left\{ \sum_{i,j \in V, i \leq j} \sum_{k,l=1}^{m_2} \theta_{ij}^{kl} \phi_{kl}(x_i, x_j) + \sum_{i \in V} \sum_{k=1}^{m_1} \theta_i^k \phi_k(x_i) \right\}. \qquad (4.64)$$

Observe that this is a finite-dimensional exponential family. Furthermore, the normalizing constant for this family will generally be intractable, requiring a d-fold integral. We choose our density estimate to be $p_{\widehat{\theta}}$, where $\widehat{\theta}$ is a solution to the score matching estimator (4.30) for this family. Furthermore, we let the number of sufficient statistics m_1, m_2 grow with the sample size n to balance the bias from truncation with the estimation error. We denote E to be the support of p^*:

$$E := \left\{ (i,j) : \|\theta_{ij}^*\| = 0 \right\}, \qquad\qquad (4.65)$$

Now, decompose the vector θ^* into the included terms and truncated terms, $\theta^* = ((\bar{\theta}^*)^\top, (\theta_T^*)^\top)^\top$ and corresponding sufficient statistics $\phi = ((\bar{\phi})^\top, (\phi_T)^\top)$. Denote $(a_T)_i(x) := \frac{\partial}{\partial x_i}(\phi_T)\cdot, i$, $A_T(x) = \text{diag}((a_T)_i(x)(a_T)_i(x)^\top)$, and $\Gamma_T^* = \mathbb{E}_p[A_T(X)]$. Applying the results

in Section 4.2.4, we have the linear relation

$$K^* = -\Gamma_T^* \theta_T^* - \Gamma^* \bar{\theta}^*. \tag{4.66}$$

In the following theorem we assume $\kappa_T := \|\Gamma_T^*\|_{\max}$ is bounded, and $\kappa_{1,\theta} = O(m_2^2)$, in addition to the assumptions for the parametric setting stated in Section 4.5.2, with the exception of Assumption 4.5.3. Since the number of statistics grows to infinity in the non-parametric case, we need more accurate accounting of the constant terms in the concentration inequality. In lieu of the concentration assumption, we have the following assumption on the boundedness of the marginals of p.

Assumption 4.5.12. For each $i, j \in V$,

$$\underline{\epsilon} \le p_{ij}(x_i, x_j) \le \bar{\epsilon}, \tag{4.67}$$

for absolute constants $\underline{\epsilon} > 0, \bar{\epsilon} < \infty$.

This assumption is mild for density estimation as it only requires bounds on the bivariate marginals rather than the full distribution. This is the same assumption used in Chapter 2 for the TRW estimator.

Theorem 4.5.13. *Suppose that the truncation parameters and regularization parameter are chosen to be*

$$m_2 \asymp n^{\frac{1}{2r+13}} \tag{4.68}$$

$$m_1 \asymp n^{\frac{1}{2r+13}} \tag{4.69}$$

$$\lambda_n \asymp \sqrt{\frac{\log nd}{n^{\frac{2r-1}{2r+13}}}} \tag{4.70}$$

and the dimension d and ρ^* satisfy

$$d = o\left(e^{n^{\frac{2r-1}{2r+13}}}\right) \tag{4.71}$$

$$\frac{1}{\rho^*} = o\left(\sqrt{\frac{\log nd}{n^{\frac{2r+1}{2r+13}}}}\right), \tag{4.72}$$

then there exists a solution $\widehat{\theta}$ such that the edge set $\widehat{E}$ satisfies

$$\mathbb{P}(\widehat{E} = E) \to 1. \tag{4.73}$$

Remark 4.5.14. If $r = 2$, and s grows as a constant, we may have

$$d = o\left(e^{n^{3/17}}\right), \tag{4.74}$$

and still ensure model selection consistency. In Chapter 2 it was shown that the sample complexity for model selection in the nonparametric pairwise model the regularized exponential series MLE using Legendre polynomials is $d = o\left(e^{n^{3/7}}\right)$, though this estimator can't be computed exactly. The optimal choice of regularization and truncation parameters is much different for these two methods. This is a consequence of different estimation errors. In our supporting lemmas (see Appendix B) we require convergence of the statistic $\widehat{\Gamma}$ to its expectation. In Appendix B we show that applying Hoeffding's inequality and a union bound,

$$\kappa_{1,\theta}\|\widehat{\Gamma} - \Gamma\|_{\max} = O_p\left(\sqrt{\frac{m_2^{12}\log nd}{n}}\right). \tag{4.75}$$

For the regularized MLE, we needed convergence of the sufficient statistics $\widehat{\mu}$, which converges at a much faster rate of $O_p\left(\sqrt{\frac{m_2^4 \log nd}{n}}\right)$. Our results agree with intuition, that the score matching statistics, derived from the derivatives of the log-density, should be harder

to estimate than the sufficient statistics.

Also it should be noted that the assumptions underlying the two results are quite different. The MLE involves conditions on the covariance of the sufficient statistics $\text{cov}_{p^*}[\phi(X)]$, while the score matching estimator requires conditions on $\Gamma^* = \mathbb{E}_{p^*}[A(X)]$. An interesting stream of future work would be to better understand the relationship between these two approaches and their assumptions.

4.6 Algorithms

In this section we consider algorithms for solving (4.30). In our experiments we denote our method QUASR, for **Qua**dratic **S**coring and **R**egularization. There are variety of generic approaches to solving problems which may be cast as the sum of a smooth convex function plus a sparsity-inducing norm [Bach et al., 2011], as well as generic solvers for solving second-order cone programs. Here we will propose two novel algorithms which exploit the unique structure of the problem at hand. First we will consider an ADMM algorithm; for a detailed exposition of this approach, see [Boyd et al., 2011]. In section 4.6.2, we consider a coordinate-wise descent algorithm for Gaussian score matching [Friedman et al., 2007].

4.6.1   Consensus ADMM

The idea behind ADMM is that the problem (4.30) can be equivalently written as

$$\min_{\theta, z} \left\{ \frac{1}{2} \sum_{i \in V} \left(\theta_{\cdot,i}^\top \widehat{\Gamma}_i \theta_{\cdot,i} + \theta_{\cdot,i}^\top \widehat{K}_{\cdot,i} \right) + \lambda \sum_{i,j \in V, i \leq j} \|z_{ij}\|_2 \right\}, \tag{4.76}$$

subject to the constraint that $\theta_{ij} = \theta_{ji} = z_{ij}$. The scaled augmented Lagrangian for this problem is given by

$$L\left(\theta, y, z\right) = \frac{1}{2} \sum_{i \in V} \left(\theta_{\cdot,i}^{\top} \widehat{\Gamma}_i \theta_{\cdot,i} + \theta_{\cdot,i}^{\top} \widehat{K}_{\cdot,i} \right) \tag{4.77}$$

$$+ \sum_{i,j \in V : i \leq j} \left(\|z_{ij}\|_2 + y_{ij}^{\top}(\theta_{ij} - z_{ij}) + y_{ji}^{\top}(\theta_{ji} - z_{ij}) \right. \tag{4.78}$$

$$\left. + \frac{\rho}{2} \|\theta_{ij} - z_{ij}\|^2 + \|\theta_{ji} - z_{ij}\|^2 \right), \tag{4.79}$$

here $\{y\} := \{y_{ij}, y_{ji}\}$ are dual variables, and ρ is a penalty parameter which we choose to be 1 for simplicity. The idea behind ADMM is to iteratively optimize L over the θ, y, z variables in turn. In the first step, since $\theta_{ij} = \theta_{ji}$ is included as a constraint and may be considered separately, L as a function of θ decouples into d independent quadratic programs, one for each "column" of θ, which may be solved in parallel. In the second step, z_{ij} pools the estimates θ_{ij} and θ_{ji} from the previous step, and applies a group shrinkage operator. The third step is a simple update of the dual variables.

Due to parallel updating of θ in step (a) and subsequent averaging in step (b), this is known as *consensus ADMM*. At convergence, the constraints $\theta_{ij} = \theta_{ji} = z_{ij}$ are binding. In practice, we stop when the average change in parameters is small:

$$\sum_{i,j \in V} \|\theta_{ij}^{(t)} - \theta_{ij}^{(t-1)}\|_1 / \sum_{i,j \in V} \|\theta_{ij}^{(t)}\|_1 < 10^{-4}. \tag{4.84}$$

In addition to parallelizing the update (a), other speedups are possible. For example, we may compute the eigenvalues Λ and eigenvectors Q of $\widehat{\Gamma}_i$, which may be computed directly from the data matrix $\left[a_i(X^1), \ldots a_i(X^n)\right]^{\top}$ using the singular value decomposition (Q being the right singular vectors, and $\sqrt{n}\Lambda$ being the squared singular values of the data matrix). We may then cache the matrix $Q(\mathrm{diag}(\Lambda + \rho)^{-1})Q^{\top}$, which is equivalent to $(\widehat{\Gamma}_i + \rho I_d)^{-1}$ up to numerical error. This can be computed for each $i \in V$, also in parallel, and only needs

1. Initialize $\theta^{(0)}$, $z^{(0)}$, $y^{(0)}$, and choose $\rho = 1$;

2. For $t = 1, \ldots$, until convergence:

   (a) Update θ for $i \in V$:

   $$\theta_{\cdot,i}^{(t)} = \left(\widehat{\Gamma}_i + \rho I_d \right)^{-1} \left(-\widehat{K}_{\cdot,i} - y_{\cdot,i}^{(t-1)} + \rho z_{\cdot,i}^{(t-1)} \right), \qquad (4.80)$$

   (b) Update z for $i, j \in V$, $i \leq j$:

   $$z_{ij}^{(t)} = \tilde{S} \left(\frac{1}{2} \left(\theta_{ij}^{(t)} + \theta_{ji}^{(t)} + y_{ij}^{(t-1)}/\rho + y_{ji}^{(t-1)}/\rho \right), \lambda/\rho \right), \qquad (4.81)$$

   where $\tilde{S}(x, \lambda) := \left(1 - \frac{\lambda}{\|x\|_2} \right)_+ x$.

   (c) Update y for $i, j \in V$, $i \leq j$:

   $$y_{ij}^{(t)} = y_{ij}^{(t-1)} + \rho \left(x_{ij}^{(t)} - z_{ij}^{(t)} \right), \qquad (4.82)$$

   $$y_{ji}^{(t)} = y_{ji}^{(t-1)} + \rho \left(x_{ji}^{(t)} - z_{ij}^{(t)} \right). \qquad (4.83)$$

Figure 4.1: QUASR Consensus ADMM

to be computed once (even if estimating over a sequence of λs). When optimizing over a path of truncation parameters m_1, m_2, one may utilize block matrix inversion formulas and the Woodbury matrix identity to avoid computing the inverse from scratch each time. In particular, let $\widehat{\Gamma}_i$ be the current matrix of statistics, and $\widehat{\Gamma}_i^{new}$ be the statistic with a higher degree of basis expansion. Then $\widehat{\Gamma}_i^{new}$ has the form for some $\widehat{b}, \widehat{C}$,

$$\widehat{\Gamma}_i^{new} = \begin{pmatrix} \widehat{\Gamma}_i & \widehat{b} \\ \widehat{b}^\top & \widehat{C} \end{pmatrix}. \tag{4.85}$$

The inverse takes the form

$$(\widehat{\Gamma}_i^{new} + \rho I)^{-1} = \begin{pmatrix} L & \\ -\left(\widehat{C} + \rho I\right)^{-1} \widehat{b}^\top L & \left(\widehat{C} + \rho I - \widehat{b}^\top \left(\widehat{\Gamma}_i + \rho I\right)^{-1} \widehat{b}\right)^{-1} \end{pmatrix}, \tag{4.86}$$

where

$$L := \left(\widehat{\Gamma}_i + \rho I - \widehat{b}^\top \left(\widehat{C} + \rho I\right)^{-1} \widehat{b}\right)^{-1} \tag{4.87}$$

$$= \left(\widehat{\Gamma}_i + \rho I\right)^{-1} - \left(\widehat{\Gamma}_i + \rho I\right)^{-1} \widehat{b}^\top \left(\widehat{C} + \rho I - \widehat{b}^\top \left(\widehat{\Gamma}_i + \rho I\right)^{-1} \widehat{b}\right)^{-1} \widehat{b} \left(\widehat{\Gamma}_i + \rho I\right)^{-1}. \tag{4.88}$$

If the dimension of $\widehat{C}$ is small relative to that of $\widehat{\Gamma}_i$, $\left(\widehat{\Gamma}_i^{new} + \rho I\right)^{-1}$ can be computed quickly using the cached $\left(\widehat{\Gamma}_i + \rho I\right)^{-1}$, without the need for any additional large matrix inversions.

4.6.2   Coordinate-wise Descent

In this section we consider a coordinate-wise descent algorithm for the Gaussian score matching problem (4.38). Coordinate-wise descent algorithms are known to be state-of-the-art for many statistical problems such as the lasso and group lasso [Friedman et al., 2007] and glasso

1. Initialize $\widehat{\Omega} = I_d$;

2. For i=1,2,...,d,1,2,..., until convergence:

   (a) for j=i,...,d:

   $$\widehat{\Omega}_{ij} \leftarrow S\left(-\frac{\left(\widehat{\Omega}_{\setminus j,i}^{\top}\widehat{\Sigma}_{\setminus j,j} + \widehat{\Omega}_{\setminus i,j}^{\top}\widehat{\Sigma}_{\setminus i,i} - 2 \cdot 1\{j = i\}\right)}{\widehat{\Sigma}_{ii} + \widehat{\Sigma}_{jj}}, \lambda\right), \qquad (4.92)$$

   and set $\widehat{\Omega}_{ji} = \widehat{\Omega}_{ij}$.

Figure 4.2: Gaussian QUASR Coordinate-wise descent

for sparse Gaussian MLE [Friedman et al., 2008]. Regularized score matching in the Gaussian case admits a particularly simple coordinate update. Consider the stationary condition for Ω in (4.38):

$$\frac{1}{2}\left(\Omega\widehat{\Sigma} + \widehat{\Sigma}\Omega\right) - I_d + \widehat{Z} = 0, \qquad (4.89)$$

where $\widehat{Z}$ is an element of the subdifferential $\partial\|\Omega\|_1$:

$$\widehat{Z}_{ij} \in \begin{cases} \{\theta_{ij} : \|\theta_{ij}\|_2 \leq 1\}, & \text{if } \|\theta_{ij}\| = 0; \\ \frac{\theta_{ij}}{\|\theta_{ij}\|_2}, & \text{if } \|\theta_{ij}\| \neq 0. \end{cases} \qquad (4.90)$$

in particular, the stationary condition for a particular Ω_{ij} is

$$\frac{1}{2}\left(\Omega_{\cdot,i}^{\top}\widehat{\Sigma}_{\cdot,j} + \widehat{\Sigma}_{\cdot,i}^{\top}\Omega_{\cdot,j}\right) - 1\{i = j\} + \widehat{Z}_{ij} = 0. \qquad (4.91)$$

Consider updating Ω_{ij} using equation (4.91), solving for Ω_{ij} and holding the other elements of Ω fixed. After some manipulation, we get a fixed point for Ω_{ij} is given by (4.92). We cycle through the entries of Ω, applying this update, and repeat until convergence.

Here $S(x, \lambda)$ is the soft thresholding function $S(x, \lambda) := \max\{|x| - \lambda, 0\}\text{sign}(x)$, and

$\backslash i := \{1, \ldots, i-1, i+1, \ldots, d\}$. Each update only requires two sparse inner products and a soft thresholding operation. As such, in our experiments this algorithm converges very quickly, sometimes much faster than glasso for the same set of data.

4.6.3   Choosing Tuning Parameters

As of yet we have not discussed how to practically choose the regularization parameter λ and for nonparametric score matching, the truncation parameters m_1, m_2. We suppose the existence of a held-out tuning set; in the absence, one may use cross-validation. If the likelihood is available, for example if fitting Gaussian score matching, or for a fixed graph which is a tree, we minimize the negative log-likelihood risk in the held out set. In the absence of the likelihood, we choose the tuning parameters to minimize the Hyvärinen score of the held out set. For a discussion on using scoring rules as a replacement for the likelihood in model selection and using *score differences* as surrogates for Bayes factors, see [Dawid and Musio, 2014].

To save on computation, we use the idea of *warm starts* which we detail in the sequel. First, observe that the first-order necessary conditions for regularized score matching are:

$$\widehat{\Gamma}\widehat{\theta} + \widehat{K} + \widehat{Z} = 0, \tag{4.93}$$

where $\widehat{Z}$ denotes the sub gradient of the regularizer $\mathcal{R}$, at $\widehat{\theta}$, which is

$$\widehat{Z}_{ij} = \begin{cases} \{x : \|x\| \leq 1\}, & \|\theta_{ij}\| = 0, \\ \frac{\theta_{ij}}{\|\theta_{ij}\|}, & o/w. \end{cases} \tag{4.94}$$

so $\widehat{\theta} = 0$ when

$$\lambda \geq \max_{ij} \|\widehat{K}_{ij}\|. \tag{4.95}$$

This allows us to choose an upper bound λ_{start} such that the solution will be the zero vector.

The idea behind warm starting is the following: we begin with estimating $\widehat{\theta}_{\lambda_{start}} = 0$. Then we fit our model on a path of λ decreasing from λ_{start}, initializing each new problem with the previous solution $\widehat{\theta}_\lambda$. The solution path for the regularized MLE is smooth as a function of λ, suggesting nearby choices of λ will provide values of $\widehat{\theta}$ which are close to one another.

We can also incorporate warm-starting in choosing m_1, m_2. For a given λ, we first estimate the model for first-order polynomials, corresponding to $m_1 = m_2 = 1$. We then increment the truncation parameters by increasing the degree of the polynomial of he sufficient statistics. We augment the previous parameter estimate vector with zeros in the place of the added parameters, and warm start *ISTA* from this vector. See Section 4.6.1 for other computation savings when augmenting the sufficient statistics when choosing m_1, m_2.

4.7 Experiments

4.7.1 Gaussian Score Matching

We begin by studying Gaussian score matching, and comparing to the regularized Gaussian MLE, using the glasso package in R [Friedman et al., 2008]. We consider experiments with two graph structures: in the first, a tree is generated randomly; this has $d - 1$ edges. In the second, a graph is generated where an edge occurs between node i and j with probability 0.1, denoted the Erdös-Renyi graph. This graph has expected number of edges $0.05 \cdot d(d-1)$. The data is scaled to have unit variance and mean zero.

Regularization Paths

Figures 4.3 and 4.4 display regularization paths for one run of these simulations, where $d = 100$ and n is either 100 or 500. Relevant variables are plotted in black. For $n = 500$, it appears that the score matching estimator does a better job screening out irrelevant variables for both graph types. For $n = 100$ they perform similarly. The score matching

estimator tends to produce nonzero parameter estimates which are larger in magnitude than the regularized MLE, which is more pronounced for $n < d$.

Risk Paths

Figures 4.5 and 4.6 show risk paths under the two graph structures; figure two has $d = 150$, with 149 included edges; figure four has $d = 100$, with 499 edges included. We choose $n = 100$, and calculate the negative log likelihood risk using a held-out dataset of size n. The plotted curves are an average of 25 simulations from the same distribution. In figure 4.5, we see the score matching estimator selects a sparser graph than the regularized MLE; furthermore, the score matching produces an estimator with smaller held-out risk. For the Erdös-Renyi simulation, the score matching estimator also selects a sparser graph, though it has risk slightly worse than the MLE.

These findings are also validated in Tables 4.1 and 4.2, varying d. For the tree graph, score matching dominates in risk for all values of d. Even for the Erdös-Renyi graph, score matching outperforms the regularized MLE when $d = 30$. Standard errors of 25 repetitions are in parentheses.

	quasr	glasso
d= 30	28.082 (0.341)	28.189 (0.319)
d= 75	72.785 (0.472)	73.202 (0.432)
d= 120	116.771 (0.578)	117.631 (0.542)
d= 150	147.26 (0.619)	148.381 (0.583)

Table 4.1: Held-out NLL error, tree graph

	quasr	glasso
d= 30	26.157 (0.428)	26.244 (0.386)
d= 50	44.351 (0.688)	44.262 (0.718)
d= 100	91.383 (1.009)	90.351 (1.077)
d= 150	139.03 (1.15)	136.793 (1.336)

Table 4.2: Held-out NLL error, ER graph
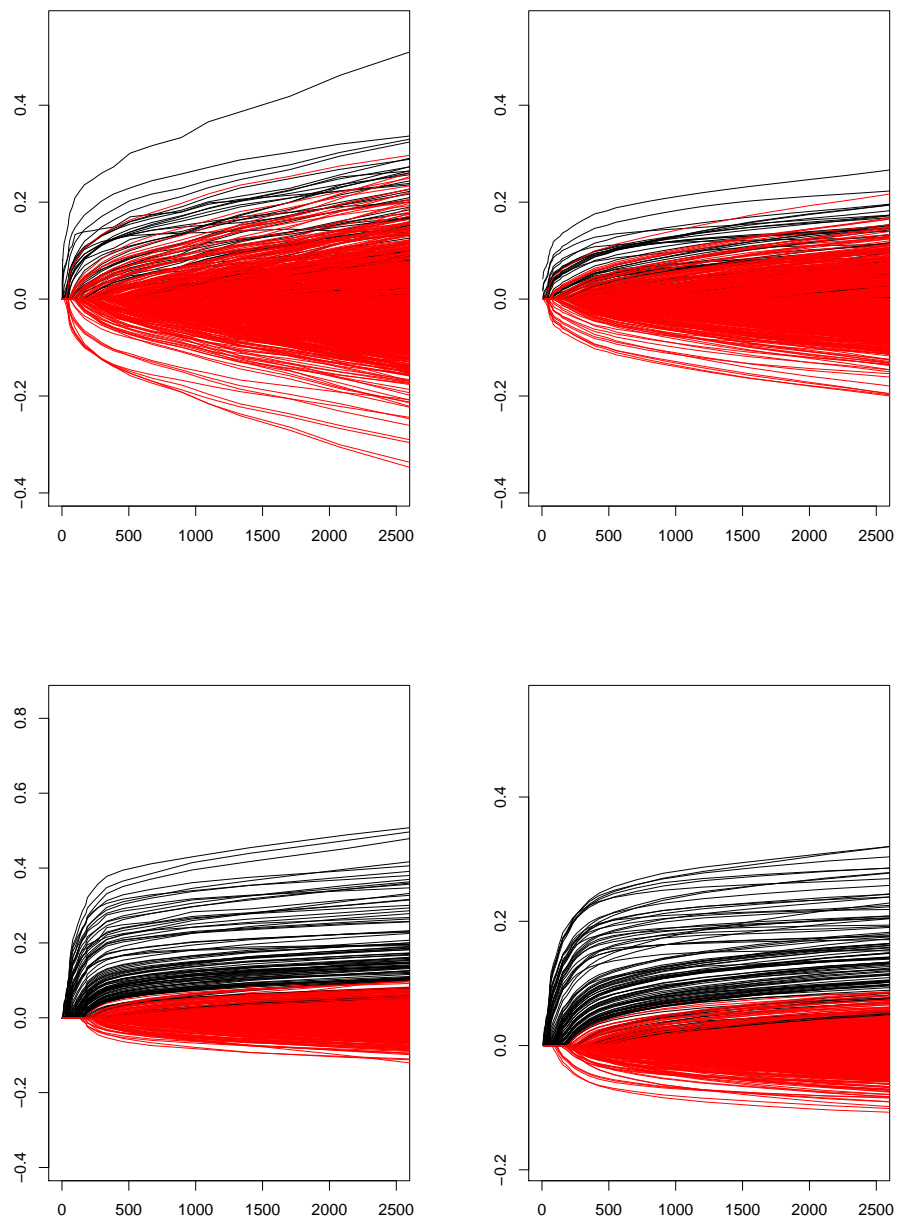
Figure 4.3: Regularization path, tree graph. d=100. Top: n=100. Bottom: n=500.
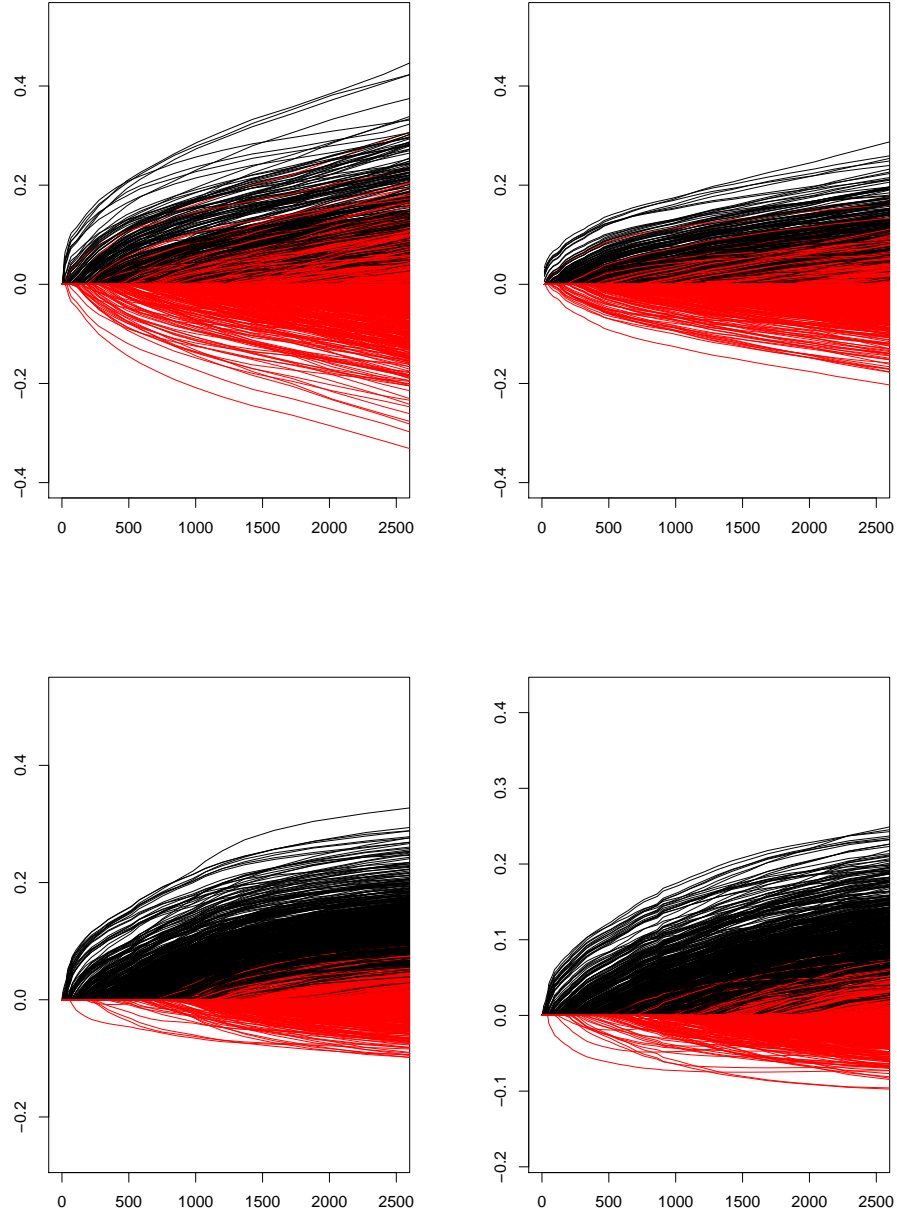
Figure 4.4: Regularization path, Erdos-Renyi graph graph. d=100. Top: n=100. Bottom: n=500.
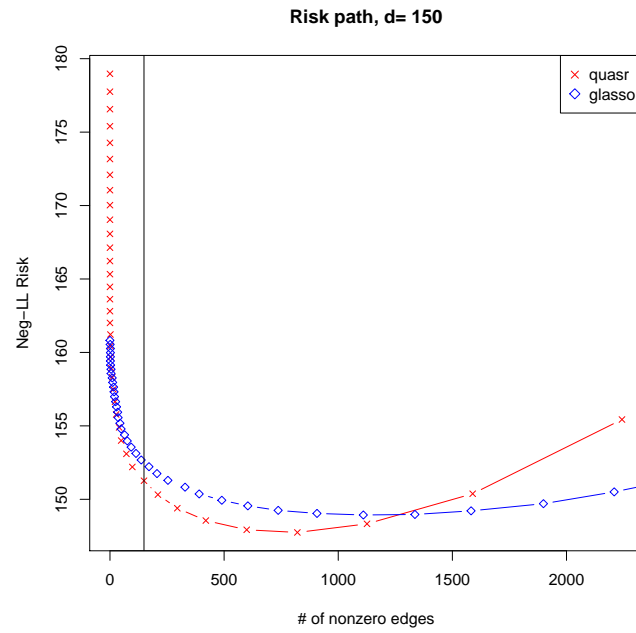
Figure 4.5: Risk path, tree graph
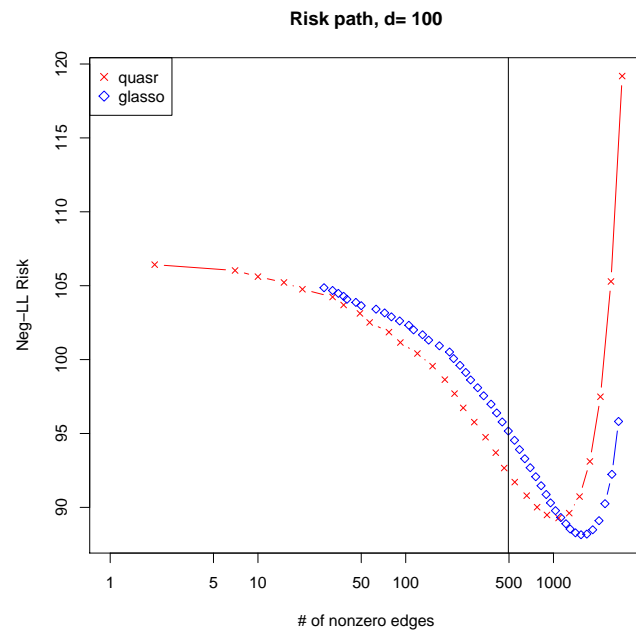
Figure 4.6: Risk path, ER graph
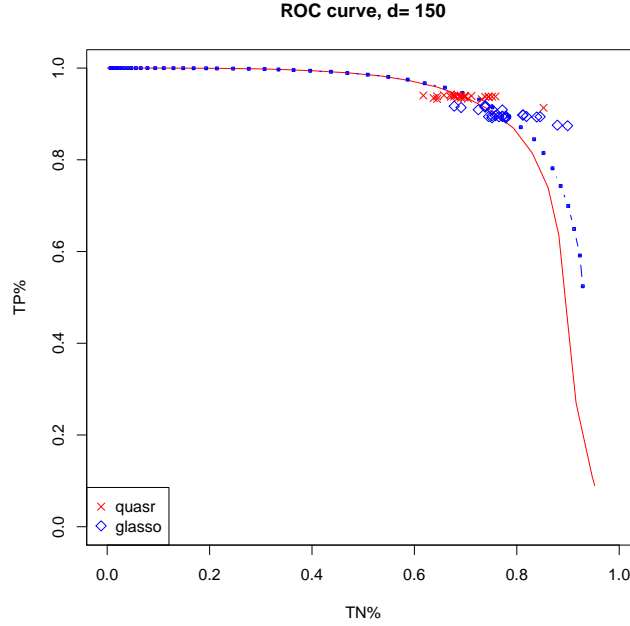
Figure 4.7: ROC curve, tree graph. Solid line: Gaussian QUASR; dotted line: glasso.

ROC Curves and Edge Selection

Figures 4.7 and 4.8 show ROC curves under the same simulation setup in the previous section. The plotted points represent the graph selected from the minimal held-out risk in each of the 25 repetitions. The two estimators display very similar ROC curves, and the score matching estimator tends to prefer higher sensitivity for lower specificity, when selecting using held-out data.

Tables 4.3 and 4.4 displays true positive and true negative rates for varing choices of d, fixing $n = 100$. The parentheses are the standard deviation for 25 repetitions of the experiment. As are suggested by the ROC curves, score matching prefers a higher sensitivity and lower specificity to the regularized MLE, and for simulations when d is relatively small compared to n, score matching has a significantly higher true positive rate with only negligible reduction in true negative rate.

ROC curve, d= 100
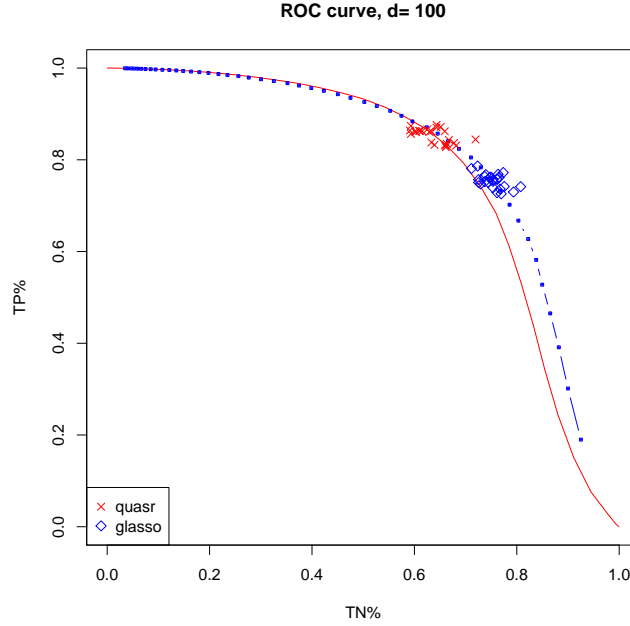
Figure 4.8: ROC curve, Erdos-Renyi graph. Solid line: Gaussian QUASR; dotted line: glasso.

Computation

In Figure 4.9 we compare the runtime of our algorithm with glasso. We simulate random Gaussian tree data with $n = 100$, $d = 50$. We fit over a path of λs and plot runtime against number of selected edges. More regularization results in sparser graphs, and so convergence is faster. In this experiment our method is much faster than glasso, sometimes by a factor of 4 or more. The gap narrows for sparse estimated graphs. This is because while our algorithm is written efficiently in $C++$, it doesn't (yet) utilize sparse matrix libraries, while glasso does. Since our coordinate-wise descent algorithm involves sparse inner products, we believe our runtimes can be improved in the sparse regime.
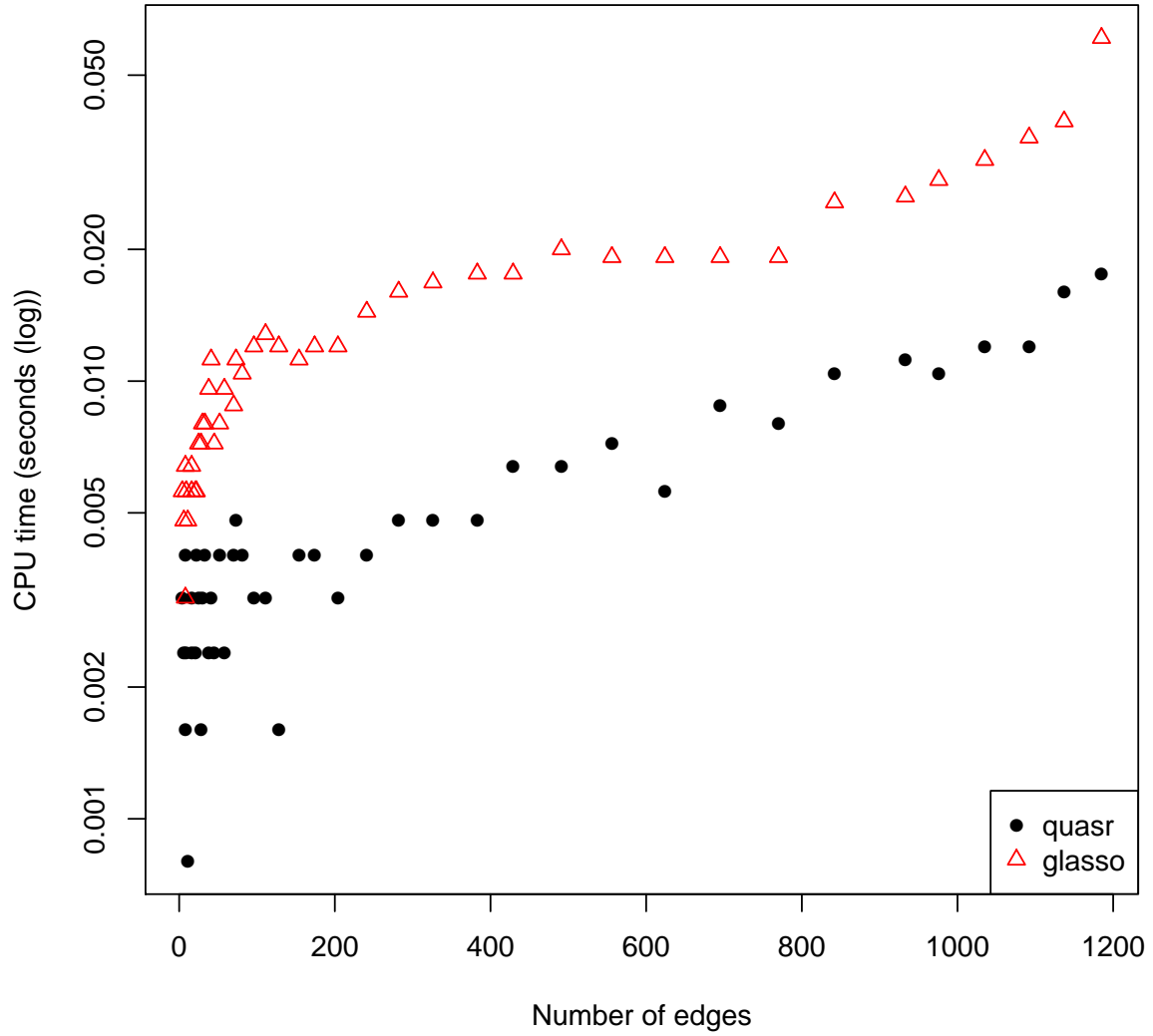
Figure 4.9: Runtime, Gaussian QUASR and glasso. Gaussian data, n=100, d=50
.

	quasr	glasso
d=30 %TN	0.941 (0.043)	0.975 (0.025)
d=30 %TP	0.778 (0.051)	0.674 (0.042)
d=75 %TN	0.821 (0.037)	0.876 (0.042)
d=75 %TP	0.892 (0.017)	0.832 (0.017)
d=120 %TN	0.751 (0.052)	0.816 (0.044)
d=120 %TP	0.922 (0.01)	0.874 (0.009)
d=150 %TN	0.699 (0.049)	0.776 (0.052)
d=150 %TP	0.936 (0.006)	0.899 (0.012)

Table 4.3: Edge selection accuracy, tree graph.

	quasr	glasso
d=30 %TN	0.969 (0.028)	0.986 (0.02)
d=30 %TP	0.743 (0.036)	0.612 (0.039)
d=50 %TN	0.878 (0.028)	0.927 (0.025)
d=50 %TP	0.785 (0.027)	0.673 (0.032)
d=100 %TN	0.633 (0.029)	0.754 (0.02)
d=100 %TP	0.858 (0.013)	0.753 (0.013)
d=150 %TN	0.486 (0.019)	0.634 (0.018)
d=150 %TP	0.892 (0.01)	0.791 (0.011)

Table 4.4: Edge selection accuracy, ER graph

4.7.2   Nonparametric Score Matching

Risk and Density Estimation

In this section we compare score matching and MLE when the sufficient statistics are chosen to be Legendre polynomials. For each choice of n, we simulate non-Gaussian data whose density factorizes as a tree. We do this by marginally applying the transformation $y = \text{sign}(x - 0.5) \mid x - 0.5 \mid^{0.6} /5 + 0.5$ to Gaussian data which has a tree factorization, which has been scaled to have means 0.5 and covariance $1/8^2$, so it fits in the unit cube; the resulting data follows a Gaussian copula distribution, which is also a pairwise distribution. We train the model using both regularized MLE and score matching under constraint that it factorizes according to the given tree, and we estimate along a path of λs and choose the regularization parameter λ to minimize the held-out risk. Since the density has a (known) tree factorization, it is possible to compute the likelihood (and hence the MLE) exactly using

86

functional message passing (Chapter 3). It is also possible to compute the marginals using the same message passing algorithm.

Figure 4.10 displays the held-out risk for both methods, with sample size varying from 24 to 5000 and $d = 20$. We average over 5 replications of the experiment. The MLE outperforms the score matching estimator, but the score matching estimators performance greatly improves relative to the MLE as n increases.

Figure 4.11 displays contours from one bivariate marginal from the aforementioned simulation. The top row shows the MLE and score matching estimator for $n = 24$, and the bottom for $n = 182$. For $n = 24$, the score matching marginal can make out much of the distinguishing features of the density such as the multiple modes, but isn't as informative as the MLE. At $n = 182$ the marginals appear almost identical. We conclude that while MLE is more efficient as may be expected, score matching performs quite well, especially with larger sample sizes. Furthermore, we emphasize that at this cost of statistical efficiency, score matching can be computed easily under any graph structure, even when the likelihood is not tractable, while MLE is typically not tractable and must be approximated.

ROC Curves

Figures 4.12 and 4.13 display ROC curves from four experiments. We simulate data with $d = 20$ and n either 30 or 100. In the first two experiments, the data is Gaussian; in the last two, it is non-Gaussian (copula). The graph is either a random spanning tree with $d-1$ edges, or a graph with each possible edge having inclusion probability 0.2, or expected number of edges $d(d-1) * .1$. The ROC curves trace the true positive and true negative rates, varying the value of λ. The curves are averaged over 10 repetitions of the experiment (with the data i.i.d. from the same distribution). We choose m_1, m_2 to maximize $\max_\lambda TP(\lambda) + TN(\lambda)$. We compare our method to the SKEPTIC estimator from [Liu et al., 2012a], which was designed in particular for model selection for copula graphical models, and the TRW estimator from Chapter 3. Our experiments show our method to do either just as well, or only slightly worse
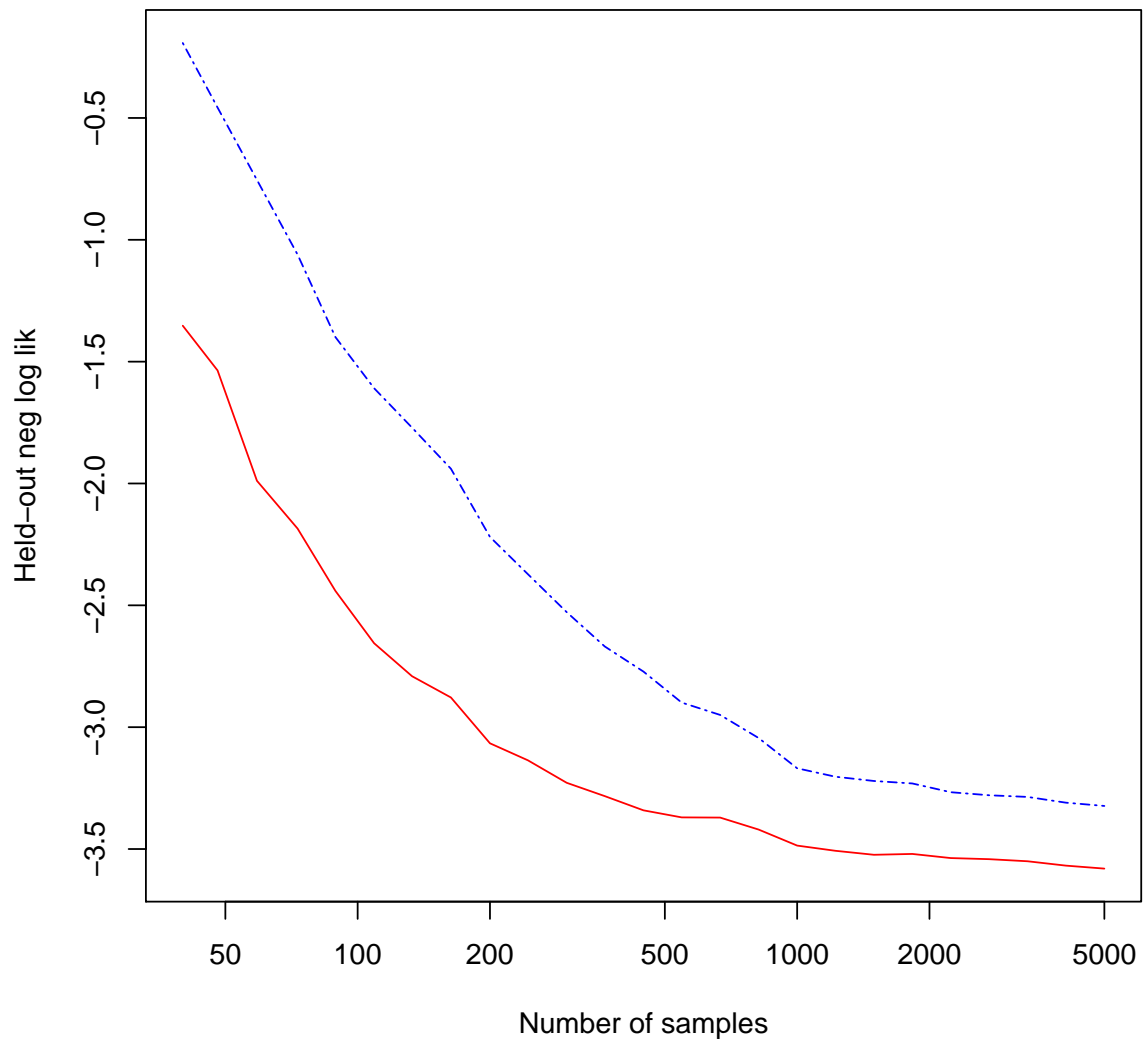
Figure 4.10: Held-out negative log likelihood for different training sizes. Red: regularized MLE; Blue dashed: score matching. Data generated from a non-Gaussian tree distribution.
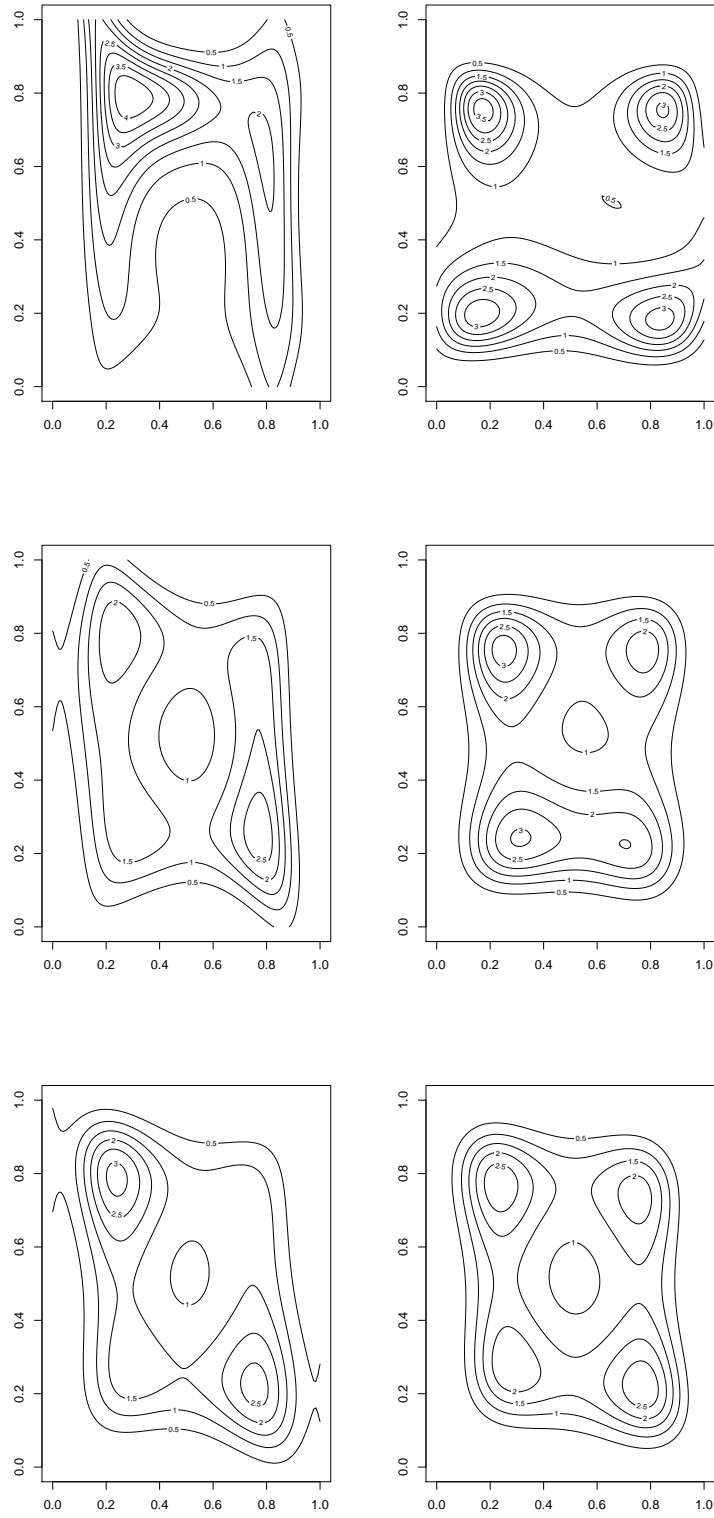
Figure 4.11: Estimated bivariate marginal from non-Gaussian tree data, d=20. Left: regularized MLE; Right: Score matching. Top: n=40. Middle: n=244 Bottom: n=1828.

than competing methods.

4.8   Discussion

This chapter introduces a new approach to estimating pairwise, continuous, exponential family graphical models. Since the normalizing constant for these models is usually intractable, we propose a new scoring rule which obviates the need for it. Our resulting estimator may be expressed as a second-order cone program. We show consistency and edge selection results for this estimator, including as special cases a new method for precision matrix estimation, and for nonparametric edge selection with exponential series. We propose algorithms for solving the convex problem which are highly scalable and amenable to parallelization. This method has good experimental performance compared to other works in the literature.
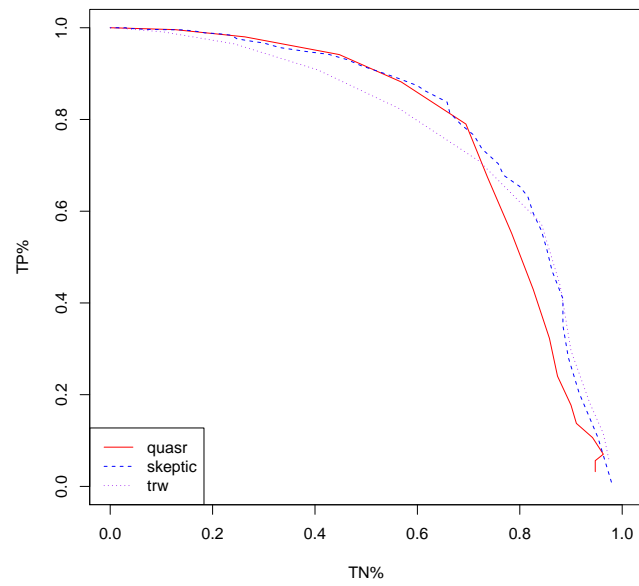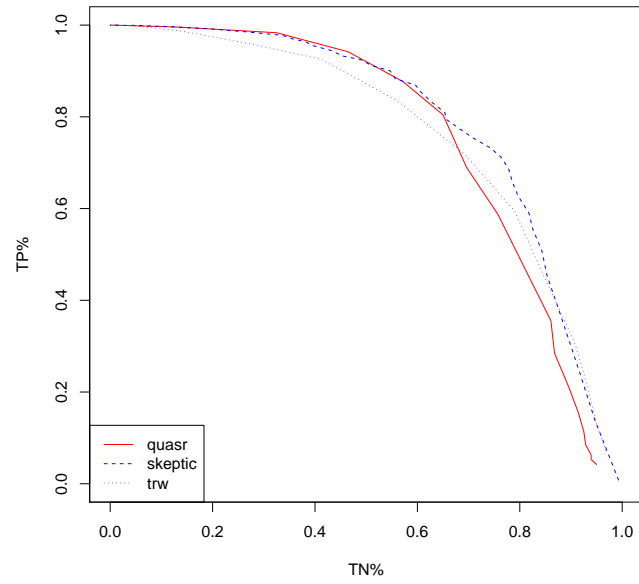
Figure 4.12: ROC Curve, Gaussian data, n=30, d=20. Top: ER graph. Bottom: Tree graph.
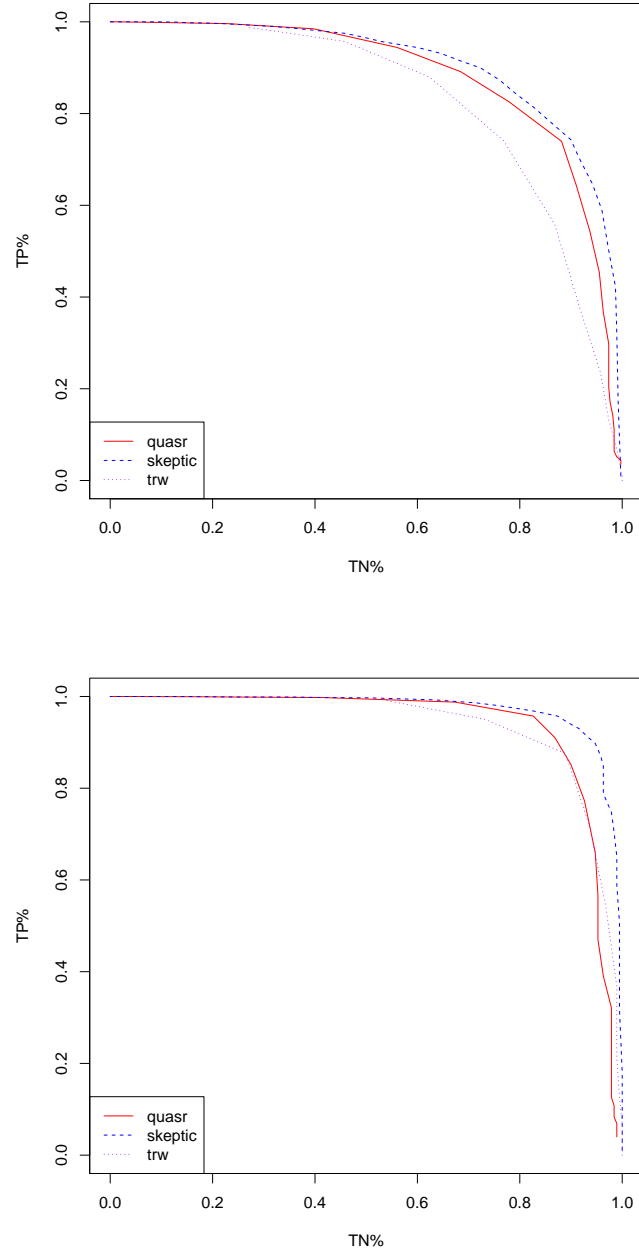
Figure 4.13: ROC Curve, non-Gaussian data, n=100, d=20. Top: ER graph. Bottom: Tree graph.

CHAPTER 5

CONCLUSIONS

This thesis proposes a new framework for density estimation, inference and structure learning for the nonparametric pairwise undirected graphical model. We consider approximating the log density using a truncated basis expansion, which results in a finite-dimensional exponential family. We consider two estimation approaches. The first is regularized maximum likelihood, for which we provide a variational approximation method. The second is a new method for estimation and graph learning of exponential families based on the scoring rule of [Hyvärinen, 2005]. We show that score matching allows for provably consistent parameter estimation and structure learning for exponential families, despite exact inference for this class of densities being intractable. As a special case we derive a new method for sparse precision matrix estimation, which performs competitively in experiments. We also derive results for the exponential series approximation, and show its performance in experiments.

This thesis contributes to two strains of literature. The first is that of pairwise density estimation [Gu, 2002; Jeon and Lin, 2006; Liu et al., 2011]. Our approach is novel in that we use exponential series estimators, rather than Mercer kernels or kernel density estimators. We also introduce a regularization approach to edge selection for learning sparse pairwise models. We show the regularized MLE estimator adapts to the unknown sparsity of the oracle pairwise density. In Chapter 3 we propose a variational approach to approximating the log-likelihood for nonparametric pairwise models; this gives an upper bound to the negative log-likelihood risk for an estimator and also gives a tractable algorithm for estimating pairwise models.

The second contribution is to the literature on high-dimensional graph selection. The literature focuses overwhelmingly on parametric models such as the Gaussian graphical model, the Ising model, or other parametric models with special structure [Yang et al., 2012]. We contribute to this literature in two ways. Firstly, our QUASR estimator works for exponential families broadly. It allows to both have consistent parameter estimation as well as edge

selection, even for exponential families which are non-normalizable. Second, applying the QUASR method to exponential series gives a method for model selection for the nonparametric pairwise graphical model. There has been recent effort to expand from parametric models to nonparametric or semiparametric graphical models, such as Gaussian copulas [Liu et al., 2012a] and forests [Liu et al., 2011]. Our approach is the first to address learning the fully nonparametric pairwise model and to demonstrate a tractable method with statistical guarantees for model selection. In addition to our theoretical and methodological contributions, we show the TRW and QUASR estimators perform well in practice, particularly compared to previous methods, and our algorithms are scalable and can take advantage of parallelization.

There still remains more work to better understand the score matching estimator and its connection to the widely-used maximum likelihood estimator. While our experiments are a start, a more detailed study comparing the assumptions of the theoretical results is warranted. It would be valuable to conduct a parallel theoretical analysis like what was done for the Danzig selector and lasso in [Bickel et al., 2009]. Also, the theoretical guarantees for the score matching approach are generally weaker than for MLE; it would be interesting to find the optimal rates for estimation and sample complexity for edge selection in the computation-limited setting (i.e., restricting to algorithms, which can be computed up so some level of accuracy in polynomial time, which excludes MLE), and given that, find an algorithm which achieves those optimal rates, if indeed the score matching approach is suboptimal.

Our work is not the final word on the subject. There are other fruitful paths for deriving good approximations or alternative estimators for exponential families with continuous potentials and for the nonparametric pairwise model, such as using semidefinite relaxations [Lasserre, 2007], or in the case of polynomial sufficient statistics, approximating the log density as a sum-of-squares polynomial, which can be normalized efficiently. It would also be of interest whether proper scoring rules [Dawid and Musio, 2014] beyond the Hyvärinen score

94

would be useful for estimating these models. In Chapter 4 we show that the nonparametric pairwise estimator of [Jeon and Lin, 2006] optimizes the *Bregman score*, which is a proper scoring rule. Exploring these connections more may give rise to more provably consistent estimators or other sound approximations or variational approaches to these problems. Also, there is more work that needs to be done to better understand the behavior of variational techniques. We hope this thesis has proven the value of the exponential series approach to nonparametric estimation, and we hope our contributions are useful tools for analysis of complex, high-dimensional data sets.

APPENDIX A

PROOFS FOR CHAPTER 2

Proof of Lemma 2.5.7. Observe that if $\lambda_n > 0$, (2.3) may alternatively be written as

$$\underset{\theta \in \mathcal{P}:\mathcal{R}^*(\theta_e) \leq C}{\operatorname{argmin}} \left\{ -\mathcal{L}(\theta) + \lambda\mathcal{R}(\theta) \right\}, \tag{A.1}$$

for some $C < \infty$. Thus the edge parameters of the solution are bounded. The only question is whether the objective approaches $-\infty$ as $\|\theta_v\|$ diverges to infinity. Examine the first order condition for θ_v:

$$\widehat{\mu}_v = \mu(\widehat{\theta})_v, \tag{A.2}$$

For any fixed $\widehat{\theta}_e$, the inverse of this equation produces a unique $\widehat{\theta}$ so long as $\widehat{\mu}_v$ belongs to the interior of the mean space, due to (5). For bases following the Haar condition, it has been shown that this is satisfied with probability one when $n > m_1$ [Crain, 1976]. Furthermore, due to the strong convexity of the objective (3), this solution will be unique when it exists. □

Let $\widehat{p} = p_{\widehat{\theta}}$, where $\widehat{\theta}$ is the solution to (2.3). Let $\tilde{\theta}$ be any solution to the population minimizer

$$\tilde{\theta} = \underset{\theta \in \tilde{\mathcal{P}}(E)}{\operatorname{argmin}} \mathrm{KL}(p \mid p_\theta). \tag{A.3}$$

$p_{\tilde{\theta}}$ is known as the *information projection* [Csiszár, 1975] and is the density satisfying $\mathbb{E}_{p_{\tilde{\theta}}}[\bar{\phi}_E] = \mathbb{E}_p[\bar{\phi}_E]$. Note that in particular, $\tilde{\theta}_{E^c} = 0$, so $p_{\tilde{\theta}}$ and $p_{\theta*}$ have the same sparsity pattern. Indeed, since $\{\bar{\phi}\}$ is a minimal family, $\tilde{\theta}$ is unique. For an information projection we have a Pythagorean theorem, which we present below.

Lemma A.0.1. [Pythagorean Theorem for KL divergence]

Suppose p_θ belongs to the finite exponential family with sufficient statistics $\{\bar{\phi}\}$. Then

$$\text{KL}(p \mid p_\theta) = \text{KL}(p \mid p_{\tilde{\theta}}) + \text{KL}(p_{\tilde{\theta}} \mid p_\theta). \tag{A.4}$$

Proof. Notice that $\log(p/p_\theta) = \log(p/p_{\tilde{\theta}}) + \log(p_{\tilde{\theta}}/p_\theta)$. Taking expectations of both sides with respect to p gives

$$\text{KL}(p \mid p_\theta) = \text{KL}(p \mid p_{\tilde{\theta}}) + \mathbb{E}_p(\log(p_{\tilde{\theta}}/p_\theta)) \tag{A.5}$$

From the moment-matching properties of the information projection, we have

$$\mathbb{E}_p(\log(p_{\tilde{\theta}}/p_\theta)) = \mathbb{E}_{p_{\tilde{\theta}}}(\log(p_{\tilde{\theta}}/p_\theta)) = \text{KL}(p_{\tilde{\theta}} \mid p_\theta). \tag{A.6}$$

The result follows. $\qquad\square$

This lemma implies that $\text{KL}(p \mid \widehat{p})$ can be split into two terms: the approximation error and estimation error. We start by analyzing the approximation error.

A.1 Approximation Error

Lemma A.1.1. [[Barron and Sheu, 1991], Lemma 1] Let p, q be two densities with respect to the measure ν. Then

$$\text{KL}(p \mid q) \leq e^{\|\log p/q - c\|_\infty} \int p \left(\log p/q - c\right)^2, \tag{A.7}$$

for any constant c.

Proof. Using the Taylor expansion of e^x, we get the bound

$$e^x - 1 - x \leq \frac{x^2}{2} e^{x+}, \tag{A.8}$$

where $x_+ = \max(0, x)$. Set $h(x) = \log p(x)/q(x) - c$. Then

$$\int p \log p/q = \int (p \log p/q + qe^c - p - c) - e^c + 1 + c \tag{A.9}$$

$$\leq \int p \left(h + e^{-h} - 1 \right) \tag{A.10}$$

$$\leq \int p \left(\frac{h^2 e^{h_+}}{2} \right) \tag{A.11}$$

$$\leq \frac{e^{\|h\|_\infty}}{2} \int p h^2. \tag{A.12}$$

$\square$

Consider the linear space $S_m := S_{m_1, m_2}$ of functions spanned by the truncated basis elements $\{\bar{\phi}\}$. Such functions need not be the logarithm of a valid density.

Theorem A.1.2. *For $f = \log p$ let*

$$\Delta_m = \|f - f_m\|_{L_2(p)}, \tag{A.13}$$

$$\gamma_m = \|f - f_m\|_\infty, \tag{A.14}$$

be the $L_2(p)$ and L_∞ approximation errors for some $f_m \in S_m$. Then the information projection $p_{\tilde{\theta}}$ satisfies

$$KL(p \mid p_{\tilde{\theta}}) \leq \frac{1}{2} e^{2\gamma_m} \Delta_m^2. \tag{A.15}$$

Thus, if γ_m is bounded,

$$KL(p \mid p_{\tilde{\theta}}) = O(\Delta_m^2). \tag{A.16}$$

Proof. Let $f_m = \langle \bar{\theta}^*, \bar{\phi} \rangle + \bar{\theta}_0^*$ be the approximation of f presumed to satisfy the given bounds on the error $f - f_m$. Define a density $p_{\bar{\theta}*}$ as $\log p_{\bar{\theta}*} = \langle \bar{\theta}^*, \bar{\phi} \rangle - Z(\bar{\theta}^*)$ and denote $f_{\bar{\theta}*} = \log p_{\bar{\theta}*}$. Using the bound from $A.1.1$ and setting $c = \theta_0^* + Z(\bar{\theta}^*)$, we have that

$$\mathrm{KL}(p \mid p_{\bar{\theta}*}) \leq \frac{e^{\|f - f_{\bar{\theta}} - \theta_0^* - Z(\bar{\theta})\|_\infty}}{2} \int p \left(f - f_{\bar{\theta}} - \theta_0^* - Z(\bar{\theta}^*) \right)^2 \tag{A.17}$$

$$\leq \frac{e^{2\|f - f_m\|_\infty}}{2} \|f - f_m\|_{L_2(p)}^2 \tag{A.18}$$

$$= \frac{e^{2\gamma_m}}{2} \Delta_m^2. \tag{A.19}$$

Since $\tilde{\theta}$ achieves the minimum KL-risk, we have

$$\mathrm{KL}(p \mid p_{\bar{\theta}}) \leq \mathrm{KL}(p \mid p_{\bar{\theta}*}) \leq \frac{e^{2\gamma_m}}{2} \Delta_m^2. \tag{A.20}$$

$\square$

Lemma A.1.3. The $L_2(p)$ and L_∞ approximations to $f = \log p$ satisfy

$$\Delta_m^2 = O\left(dm_1^{-2r} + |E| \, m_2^{-2r} \right), \tag{A.21}$$

$$\gamma_m = O\left(dm_1^{-r+\alpha+1/2} + |E| \, m_2^{-r+2\alpha+1/2} \right), \tag{A.22}$$

and the truncated variables θ_T^* satisfy

$$\|\theta_T^*\|_\infty = o(\max\{m_1^{-r-1/2}, m_2^{-r-1/2}\}). \tag{A.23}$$

Proof. By the boundedness of p_{ij} and Parceval's identity,

$$\Delta_m^2 = \int p \langle \theta_T^*, \phi_V \rangle^2 \tag{A.24}$$

$$\leq \bar{\epsilon} \int \langle \theta_T^*, \phi_V \rangle^2 \tag{A.25}$$

$$= \bar{\epsilon} \|\theta_T^*\|_2^2. \tag{A.26}$$

Observe that

$$\|\theta_T^*\|^2 = \sum_{i \in V} \sum_{k > m_1} (\theta_i^k)^2 \tag{A.27}$$

$$+ \sum_{(i,j) \in E} \left(\sum_{k \geq 1, l > m_2} (\theta_{ij}^{kl})^2 + \sum_{k > m_2, l \leq m_2} (\theta_{ij}^{kl})^2 \right)$$

Now, we have for each $i \in V$,

$$\sum_{k > m_1} (\theta_i^k)^2 = \sum_{k > m_1} (\theta_i^k)^2 k^{2r} k^{-2r} \tag{A.28}$$

$$\leq \left(\sum_{k > m_1} (\theta_i^k)^2 k^{2r} \right) m_1^{-2r} \tag{A.29}$$

$$\leq C m_1^{-2r}, \tag{A.30}$$

and also

$$\sum_{k \geq 1, l \geq m_2} (\theta_{ij}^{kl})^2 = \sum_{k \geq 1, l \geq m_2} (\theta_{ij}^{kl})^2 k^{2r_i} l^{2r_j} k^{-2r_i} l^{-2r_j} \tag{A.31}$$

$$\leq \left(\sum_{k \geq 1, l \geq m_2} (\theta_{ij}^{kl})^2 k^{2r_i} l^{2r_j} \right) m_2^{-2r_j} \tag{A.32}$$

$$= O(m_2^{-2r}). \tag{A.33}$$

We similarly get a bound of $O(m_2^{-2r})$ when summing over $k > m_2, l \leq m_2$. Combining we have

$$\Delta_m^2 = O(\|\theta_T^*\|_2^2) = O(d m_1^{-2r} + |E| m_2^{-2r}). \tag{A.34}$$

To analyze γ_m, observe that

$$\gamma_m = \left| \sum_{i \in V} \sum_{k > m_1} \theta_i^k \phi_k \right. \tag{A.35}$$

$$\left. + \sum_{(i,j) \in E} \left(\sum_{k \geq 1, l > m_2} (\theta_{ij}^{kl}) \phi_{kl} + \sum_{k > m_2, l \leq m_2} (\theta_{ij}^{kl}) \phi_{kl} \right) \right|$$

$$\leq \sum_{i \in V} \sum_{k > m_1} \left| \theta_i^k b(k) \right| \tag{A.36}$$

$$+ \sum_{(i,j) \in E} \left(\sum_{k \geq 1, l > m_2} \left| (\theta_{ij}^{kl}) b(k) b(l) \right| + \sum_{k > m_2, l \leq m_2} \left| (\theta_{ij}^{kl}) b(k) b(l) \right| \right).$$

Then we have

$$\sum_{k > m_1} \left| \theta_i^k b(k) \right| = O \left(\sum_{k > m_1} \left| \theta_i^k k^\alpha \right| \right) \tag{A.37}$$

$$\leq O \left(\left(\sum_{k > m_1} (\theta_i^k k^r)^2 \right)^{1/2} \left(\sum_{k > m_1} k^{2(-r+\alpha)} \right)^{1/2} \right) \tag{A.38}$$

$$= O \left(m_1^{-r+\alpha+1/2} \right). \tag{A.39}$$

Similarly,

$$\sum_{k, l > m_2} \left| \theta_{ij}^{kl} b(k) b(l) \right| = O \left(\sum_{k \geq 1, l > m_2} \left| \theta_{ij}^{kl} k^\alpha l^\alpha \right| \right) \tag{A.40}$$

$$\leq O \left(\left(\sum_{k \geq 1, l > m_2} (\theta_{ij}^{kl} k^{r_i} l^{r_j})^2 \right)^{1/2} \left(\sum_{k \geq 1, l > m_2} k^{2(-r_i+\alpha)} l^{2(-r_j+\alpha)} \right)^{1/2} \right) \tag{A.41}$$

$$= O \left(m_2^{-r_j+1/2+\alpha} \right). \tag{A.42}$$

The result is similar summing over $k \geq m_2, l \leq m_2$. Combining, we get

$$\gamma_m = O\left(dm_1^{-r+\alpha+1/2} + |E| \, m_2^{-r+\alpha+1/2}\right) \tag{A.43}$$

Finally, the series $\sum_{k\geq 1}(\theta_i^k)^2 k^{2r}$ converges if and only if $\left|(\theta_i^k)^2\right| = o(k^{-2r-1})$. Similarly, $\sum_{k\geq 1}(\theta_{ij}^{kl})^2 k^{2r}$ converges if and only if $\left|(\theta_{ij}^{kl})^2\right| = o(k^{-2r-1})$, and similarly when summing over l. It follows that

$$\|\theta_T^*\|_\infty = o(\max\{m_1^{-r-1/2}, m_2^{-r-1/2}\}). \tag{A.44}$$

$\square$

A.2 Estimation Error

Consider the Taylor expansion of $Z(\theta)$ up to order 2, noting that $\nabla Z(\theta) = \mathbb{E}_\theta[\bar{\phi}] = \mu(\theta)$ and $\nabla^2 Z(\theta) = \text{cov}_{p_\theta}[\bar{\phi}]$:

$$Z(\theta + \delta) = Z(\theta) + \langle \delta, \mu(\theta) \rangle + \frac{1}{2}\delta^\top \left(\text{cov}_{p_{\theta+z\delta}}[\bar{\phi}]\right) \delta, \tag{A.45}$$

for some $z \in [0, 1]$. See for example [Kakade et al., 2010; Portnoy, 1988].

Consider the function

$$\mathcal{E}(\delta) = \mathcal{L}(\tilde{\theta} + \delta) - \mathcal{L}(\tilde{\theta}) - \lambda(R(\tilde{\theta} + \delta) - R(\tilde{\theta})).$$

Note that $\mathcal{E}(0) = 0$, and so $\widehat{\delta} = \widehat{\theta} - \tilde{\theta}$ must satisfy $E(\widehat{\delta}) \geq 0$, if $\widehat{\theta}$ exists.

Lemma A.2.1. Denote $\mu^* = \mathbb{E}_p[\bar{\phi}] = \mathbb{E}_{\tilde{\theta}}[\bar{\phi}]$. Suppose $\lambda_n \geq \mathcal{R}^*(\widehat{\mu}_e - \mu_e^*)$, $\|\widehat{\mu}_v - \mu_v\| \leq \frac{r_{\mathcal{S}}c_{\mathcal{S}}}{\sqrt{2}}$

and $\lambda\sqrt{\tilde{E}} \leq \frac{\sqrt{2}r_\mathcal{S}c_\mathcal{S}}{3}$, then the regularized MLE $\widehat{\theta}_\lambda$ is unique and satisfies

$$\|\widehat{\theta}_v - \tilde{\theta}_v\| \leq \frac{1}{2c_\mathcal{S}}\|\widehat{\mu}_v - \bar{\mu}_v^*\|, \tag{A.46}$$

$$\|\widehat{\theta}_e - \tilde{\theta}_e\| \leq \frac{3}{4c_\mathcal{S}}\lambda_n\sqrt{|E|}, \tag{A.47}$$

$$\mathrm{KL}(p_{\tilde{\theta}} \mid p_{\widehat{\theta}_\lambda}) \leq \frac{1}{2c_\mathcal{S}}\left(\|\widehat{\mu}_v - \mu_v^*\|^2 + \frac{9}{4}\lambda^2\,|E|\right). \tag{A.48}$$

Proof. Using the cumulant expansion of Z, we have for some $z \in [0, 1]$,

$$\mathcal{L}(\tilde{\theta} + \delta) - \mathcal{L}(\tilde{\theta}) = \langle \delta, \widehat{\mu} \rangle + Z(\tilde{\theta}) - Z(\tilde{\theta} + \delta) \tag{A.49}$$

$$= \langle \delta, \widehat{\mu} - \mu^* \rangle - \frac{1}{2}\delta^\top\left(\mathrm{cov}_{p_{\tilde{\theta}+z\delta}}[\bar{\phi}]\right)\delta. \tag{A.50}$$

Consider the set $\mathcal{S} := \{\theta : \|\theta - \theta^*\| \leq r_\mathcal{S}\}$ for a constant $r_\mathcal{S}$. $\mathcal{S}$ is a convex and compact set; since $\bar{\phi}$ is minimal, $\nabla^2 Z(\theta) = \mathrm{cov}_\theta[\bar{\phi}] > 0$ (3) for all $\theta \in \mathcal{S}$; thus $-\mathcal{L}$ is strongly convex over $\mathcal{S}$ with $\delta^\top(\mathrm{cov}_\theta[\bar{\phi}])\delta > c_\mathcal{S}\|\delta\|^2$ for a constant $c_\mathcal{S}$.

By the (generalized) Cauchy-Schwarz inequality,

$$\left|\langle \mu_e - \widehat{\mu}_e, \delta_e \rangle\right| \leq \mathcal{R}(\delta_e)\mathcal{R}^*(\mu_e - \widehat{\mu}_e) \tag{A.51}$$

$$\leq \frac{\lambda_n}{2}\mathcal{R}(\delta_e) \tag{A.52}$$

$$= \frac{\lambda_n}{2}\left(\mathcal{R}(\delta_{\tilde{\mathcal{P}}}) + \mathcal{R}(\delta_{\tilde{\mathcal{P}}^\perp})\right). \tag{A.53}$$

the last line following because $\mathcal{R}$ is decomposable with respect to $\tilde{\mathcal{P}}$. Further, from [Negahban et al., 2012] Lemma 3, because $\mathcal{R}$ is decomposable, it holds that

$$\mathcal{R}(\tilde{\theta} + \delta) - \mathcal{R}(\tilde{\theta}) \geq \mathcal{R}(\delta_{\tilde{\mathcal{P}}^\perp}) - \mathcal{R}(\delta_{\tilde{\mathcal{P}}}). \tag{A.54}$$

Combining (B.17) and (B.16),

$$\langle \delta_e, \widehat{\mu}_e - \mu_e \rangle - \frac{\lambda_n}{2}\left(\mathcal{R}(\tilde{\theta} + \delta) - \mathcal{R}(\tilde{\theta})\right) \leq \frac{3\lambda_n}{2}\mathcal{R}(\delta_{\widetilde{\mathcal{P}}}) - \frac{\lambda_n}{2}\mathcal{R}(\delta_{\widetilde{\mathcal{P}}\perp}). \tag{A.55}$$

Using the subspace compatibility constant we have that

$$\mathcal{R}(\delta_{\widetilde{\mathcal{P}}}) \leq \sqrt{|E|}\|\delta_{\widetilde{\mathcal{P}}}\| \leq \sqrt{|E|}\|\delta_e\|. \tag{A.56}$$

Additionally, the second term in (B.19) is negative and can be ignored. From Cauchy-Schwarz we also have

$$\langle \delta_v, \mu_v - \widehat{\mu}_v \rangle \leq \|\mu_v - \widehat{\mu}_v\|\|\delta_v\|. \tag{A.57}$$

Now, consider the set

$$\mathcal{C} = \left\{ \delta : \|\delta_v\| \leq \frac{1}{2c_{\mathcal{S}}}\|\widehat{\mu}_v - \mu_v\|; \|\delta_e\| \leq \frac{3}{4c_{\mathcal{S}}}\lambda\sqrt{\tilde{E}} \right\} \tag{A.58}$$

denote the boundary of $\mathcal{C}$ by $\partial \mathcal{C}$ and its interior by int$\mathcal{C}$. Note that $0 \in \mathcal{C}$. If $E(\delta) < 0$ for each $\delta \in \partial \mathcal{S}$, noting that $\mathcal{E}(0) = 0$, by the convexity of $\mathcal{E}$ and the fact that $\mathcal{C}$ is a compact set, it must be that $\widehat{\delta} \in \text{int}\mathcal{C}$. Suppose that $\|\widehat{\mu}_v - \mu_v^*\| \leq \frac{2r_{\mathcal{S}}c_{\mathcal{S}}}{\sqrt{2}}$ and $\frac{3}{2}\lambda\sqrt{|E|} \leq \frac{2r_{\mathcal{S}}c_{\mathcal{S}}}{\sqrt{2}}$, so that for $\delta \in \mathcal{C}$, $\|\delta\| \leq r_{\mathcal{S}}$, so that for any $z \in [0,1]$, $\tilde{\theta} + z\delta \in \mathcal{S}$. Using (A.50) and (B.19) we

obtain for $\delta \in \partial \mathcal{C}$,

$$\mathcal{E}(\delta) = \mathcal{L}(\tilde{\theta} + \delta) - \mathcal{L}(\tilde{\theta}) - \lambda \left(\mathcal{R}(\tilde{\theta} + \delta) - \mathcal{R}(\tilde{\theta}) \right) \tag{A.59}$$

$$\leq \langle \delta_v, \widehat{\mu}_v - \mu_v^* \rangle + \langle \delta_e, \widehat{\mu}_e - \mu_e^* \rangle - c_{\mathcal{S}} \|\delta\|^2 - \lambda \left(\mathcal{R}(\tilde{\theta} + \delta) - \mathcal{R}(\tilde{\theta}) \right) \tag{A.60}$$

$$\leq \|\delta_v\| \|\widehat{\mu}_v - \mu_v^*\| + \frac{3\lambda\sqrt{\tilde{E}}}{2} \|\delta_e\| - c_{\mathcal{S}} \|\delta\|^2 \tag{A.61}$$

$$\leq \|\delta_v\| \left(\|\widehat{\mu}_v - \mu_v^*\| - c_{\mathcal{S}} \|\delta_v\| \right) + \|\delta_e\| \left(\frac{3\lambda\sqrt{\tilde{E}}}{2} - c_{\mathcal{S}} \|\delta_e\| \right) \tag{A.62}$$

$$< 0. \tag{A.63}$$

The claim is verified.

Furthermore, since

$$\mathrm{KL}(p_{\tilde{\theta}} \mid p_{\widehat{\theta}}) = \langle \tilde{\theta} - \widehat{\theta}, \mu^* \rangle - Z(\tilde{\theta}) + Z(\widehat{\theta}), \tag{A.64}$$

note that

$$\mathcal{E}(\widehat{\delta}) = \langle \widehat{\delta}, \widehat{\mu} \rangle - Z(\tilde{\theta} + \widehat{\delta}) + Z(\tilde{\theta}) - \lambda(\mathcal{R}(\tilde{\theta} + \delta) - \mathcal{R}(\tilde{\theta})) \tag{A.65}$$

$$= -\mathrm{KL}(p_{\tilde{\theta}} \mid p_{\widehat{\theta}}) + \langle \widehat{\delta}, \widehat{\mu} - \mu^* \rangle - \lambda(\mathcal{R}(\tilde{\theta} + \delta) - \mathcal{R}(\theta^*)). \tag{A.66}$$

because $\mathcal{E}(\widehat{\delta}) \geq 0$,

$$\mathrm{KL}(p_{\tilde{\theta}} \mid p_{\widehat{\theta}}) \leq \langle \widehat{\delta}, \widehat{\mu} - \mu^* \rangle - \lambda \left(\mathcal{R}(\tilde{\theta} + \widehat{\delta}) - \mathcal{R}(\tilde{\theta}) \right) \tag{A.67}$$

$$\leq \|\widehat{\delta}_v\| \|\widehat{\mu}_v - \mu_v^*\| + \frac{3\lambda\sqrt{|E|}}{2} \|\widehat{\delta}_e\| \tag{A.68}$$

$$\leq \frac{1}{2c_{\mathcal{S}}} \left(\|\widehat{\mu}_v - \mu_v^*\|^2 + \frac{9\lambda^2 \tilde{E}}{4} \right). \tag{A.69}$$

$\square$

Lemma A.2.2. Suppose that the univariate sufficient statistics satisfy $\mid \phi_k \mid \leq b(k)$ for all k and $b(k)$ is increasing in k. For any $t > 0$,

$$\mathbb{P}(\|\widehat{\mu}_v - \mu_v\|^2 > t) \leq \frac{\bar{\epsilon} d m_1}{nt}, \tag{A.70}$$

$$\mathbb{P}(\mathcal{R}^*(\widehat{\mu}_e - \mu_e^*) > t) \leq 2 \exp \left\{ \log(d^2 m_2^2) - \frac{nt^2}{8\bar{\epsilon} m_2^2 b(m_2)^4} \right\}, \tag{A.71}$$

$$\mathbb{P}(\|\widehat{\mu} - \mu^*\|_\infty > t) \leq 2 \exp \left\{ \log(dm_1 + d^2 m_2) - \frac{nt^2}{8\bar{\epsilon} \max\{b(m_2)^4, b(m_1)^2\}} \right\}. \tag{A.72}$$

Proof. Let $(\mathrm{cov}_p[\bar{\phi}])_v$ be the covariance matrix of the sufficient statistics ϕ restricted to the vertex parameters. Then $((\mathrm{cov}_p[\bar{\phi}])_v)^{-1/2}(\widehat{\mu}_v - \mu_v)$ is a vector of dm_1 mean zero, uncorrelated random variables each with variance $1/n$. It follows from Markov's inequality that

$$\mathbb{P}\left((\widehat{\mu}_v - \mu_v)^\top ((\mathrm{cov}_p[\phi])_v)^{-1}(\widehat{\mu}_v - \mu_v) > t \right) \leq \frac{dm_1}{nt}. \tag{A.73}$$

From the boundedness of the marginals p_{ij} and the orthonormality of $\{\phi\}$, $\bar{\epsilon}^{-1}\|\delta\|^2 \leq \delta^\top (\mathrm{cov}_p[\phi])^{-1}\delta$ for any δ. It follows that

$$\mathbb{P}(\|\widehat{\mu}_v - \mu_v\|^2 > \bar{\epsilon} t) \leq \frac{dm_1}{nt}. \tag{A.74}$$

Now, to bound $\mathcal{R}^*(\widehat{\mu}_e - \mu_e^*)$ we will need to find an exponential concentration for each $\|\widehat{\mu}_{ij} - \mu_{ij}\|$. Suppose that the univariate sufficient statistic is bounded: $\mid \phi_k \mid \leq b(k)$ for all k. Applying Hoeffding's inequality, we get for each indices k, l,

$$\mathbb{P}(\bar{\epsilon}^{-1/2} |\widehat{\mu}_{ij}^{kl} - \mu_{ij}^{kl}| \geq t) \leq 2 \exp \left\{ -n \frac{t^2}{8b(k)^2 b(l)^2} \right\}, \tag{A.75}$$

and so

$$\mathbb{P}\left(\bar{\epsilon}^{-1}\|\widehat{\mu}_{ij} - \mu_{ij}\|^2 \geq m_2^2 b(m_2)^4 t^2\right) \leq \mathbb{P}\left(\bigcup_{k,l}\left\{\bar{\epsilon}^{-1/2} \mid \widehat{\mu}_{ij}^{kl} - \mu_{ij}^{kl} \mid \geq tb(k)b(l)\right\}\right) \quad \text{(A.76)}$$

$$\leq \sum_{k,l=1}^{m_2} \exp\left\{-\frac{nt^2}{8}\right\} \quad \text{(A.77)}$$

$$\leq 2m_2^2 \exp\left\{-\frac{nt^2}{8}\right\}. \quad \text{(A.78)}$$

Applying the union bound once more,

$$\mathbb{P}(\mathcal{R}^*(\widehat{\mu}_e - \mu_e^*) \geq t^2\bar{\epsilon}) \leq 2\exp\left\{\log(d^2 m_2^2) - \frac{nt^2}{8m_2^2 b(m_2)^4}\right\}. \quad \text{(A.79)}$$

$\square$

Proof of Theorem 2.5.8. From theorem A.1.2, we have the approximation error

$$\mathrm{KL}(p \mid p_\theta) = O\left(e^{2\gamma_m}\Delta_m^2\right), \quad \text{(A.80)}$$

From A.1.3, we know that

$$\Delta_m^2 = O(dm_1^{-2r} + |E|^{-2r}), \quad \text{(A.81)}$$

and

$$\gamma_m = O(dm_1^{-r+\alpha+1/2} + |E|\, m_2^{-r+\alpha+1/2}). \quad \text{(A.82)}$$

To keep e^{γ_m} bounded, we require

$$m_1 = \Omega(d^{\frac{1}{r-\alpha-1/2}}), \tag{A.83}$$

$$m_2 = \Omega(|E|^{\frac{1}{r-\alpha-1/2}}). \tag{A.84}$$

From lemma A.2.1, the estimation error satisfies, for $\lambda_n \geq \mathcal{R}^*(\widehat{\mu}_e - \mu_e^*)$,

$$\mathrm{KL}(p_{\tilde{\theta}} \mid p_{\widehat{\theta}}) = O\left(\lambda_n^2 |E| + \|\widehat{\mu}_v - \mu_v^*\|_2^2 d\right). \tag{A.85}$$

Applying lemma A.2.2, $\mathcal{R}^*(\widehat{\mu}_e - \mu_e^*) > \lambda_n$ with probability no more than

$$2\exp\left\{\log(m_2^2 d^2) - \frac{\lambda_n^2 n}{8\bar{\epsilon}m_2^2 b(m_2)^4}\right\}, \tag{A.86}$$

choosing $\lambda_n = \frac{2\bar{\epsilon}^{1/2}m_2 b(m_2)^2 \log^{1/2}(md)}{n^{1/2}}$, the probability is bounded by $\frac{2}{m_2^2 d^2}$. Also,

$$\|\widehat{\mu}_v - \mu_v^*\|_2^2 > \frac{d\bar{\epsilon}m_1}{n}t \tag{A.87}$$

with probability no more than $\frac{1}{t}$. Thus applying lemma A.0.1, under the stated conditions we have

$$\mathrm{KL}(p \mid p_{\widehat{\theta}}) = \mathrm{KL}(p \mid p_{\tilde{\theta}}) + \mathrm{KL}(p_{\tilde{\theta}} \mid p_{\widehat{\theta}}) \tag{A.88}$$

$$= O_p\left(dm_1^{-2r} + |E| m_2^{-2r} + \frac{m_2^2 b(m_2)^4 \log(nd)}{n} |E| + \frac{m_1}{n}d\right) \tag{A.89}$$

$\square$

A.3   Model Selection

Taylor expansion in Hilbert Space

For any $f = \langle \theta, \phi \rangle \in W_r^2$, $\|\theta\|_2 = \|f\|_{L_2} < \infty$. Thus the set of such θ belong to a Hilbert space. Applying the Taylor expansion with remainder for Hilbert spaces, for $\|\delta\| < \infty$ we have

$$Z(\theta + \delta) = Z(\theta) + DZ(\theta) \cdot \delta + \frac{1}{2} D^2 Z(\delta) \cdot \delta^2 + \frac{1}{6} D^3(\theta + z\delta) \cdot \delta^3. \tag{A.90}$$

here z is some point in $[0, 1]$, and $D^k Z$ is the kth Fréchet derivative of Z represented as a multilinear map, $\cdot \delta^k$ denotes evaluation at $(\delta, \ldots, \delta)$. Now, consider θ^*, δ which can be decomposed to a truncated vector and a remainder, $\theta^* = (\bar{\theta}^*, \theta_T^*)^\top$ and $\delta = (\bar{\delta}, \delta_T)^\top$ and write $\phi = (\bar{\phi}, \phi_T)$. Taking the derivative of Z with respect to $\bar{\theta}^*$, and inputting the particular expression for the terms of the Taylor series (see [Portnoy, 1988]), we get the Taylor expansion for $\bar{\mu}$,

$$\bar{\mu}(\theta^* + \delta) - \bar{\mu}(\theta^*) = \text{cov}_{\theta^*}[\bar{\phi}, \phi]\delta + \frac{1}{2}\mathbb{E}_{\theta^* + z\delta}[\langle \delta, \phi \rangle^2 \bar{\phi}] \tag{A.91}$$

$$= \text{cov}_{\theta^*}[\bar{\phi}, \bar{\phi}]\bar{\delta} + \text{cov}_{\theta^*}[\bar{\phi}, \phi_T] \cdot \delta_T \tag{A.92}$$

$$+ \underbrace{\frac{1}{2}\mathbb{E}_{\theta^\dagger}\left[\left(\langle \bar{\delta}, \bar{\phi} - \mathbb{E}_{\theta^\dagger}[\bar{\phi}]\rangle + \langle \delta_T, \phi_T - \mathbb{E}_{\theta^\dagger}[\phi_T]\rangle\right)^2 \cdot (\bar{\phi} - \mathbb{E}_{\theta^\dagger}[\bar{\phi}])\right]}_{\bar{R}}.$$

Here $\theta^\dagger = \theta^* + z\delta$ for some $z \in [0, 1]$. We denote the final term the *remainder*, $\bar{R} := \bar{R}(\delta)$.

Primal Dual Witness

Our proof technique is the *primal dual witness* method, used previously in analysis of model selection for graphical models [Jason D. Lee, 2014; Ravikumar et al., 2011]. It proceeds as follows: construct a primal-dual pair $(\widehat{\theta}, \widehat{Z})$ which satisfies $\text{supp}(\widehat{\theta}) = \text{supp}(\theta^*)$, and also

satisfies the stationary conditions for 2.3 with high probability. The stationary conditions for 2.3 are

$$\widehat{\mu} - \mu(\widehat{\theta}) + \lambda_n \widehat{Z} = 0, \tag{A.93}$$

where $\widehat{Z}$ is an element of the subdifferential $\partial \mathcal{R}(\widehat{\theta}_e)$:

$$(\partial \mathcal{R}(\theta_e))_{ij} = \begin{cases} \{\theta_{ij} : \|\theta_{ij}\|_2 \leq 1\}, & \text{if } \theta_{ij} = 0; \\ \frac{\theta_{ij}}{\|\theta_{ij}\|} \|_2, & \text{if } \theta_{ij} \neq 0; \\ 0, & \text{if } i = j. \end{cases} \tag{A.94}$$

From this we may conclude that the solution to 2.3 is sparsistent. In particular, we have the following steps:

1. Set $\widehat{\theta}_{E^c} = 0$;

2. Set $\widehat{Z}_{ij} = \partial \mathcal{R}(\theta_e^*)_{ij} = \frac{\bar{\theta}_{ij}^*}{\|\theta_{ij}^*\|}$ for $(i,j) \in E$;

3. Given these choices for $\widehat{\theta}_{E^c}$ and $\widehat{Z}_E$, choose $\widehat{\theta}_E$ and $\widehat{Z}_{E^c}$ to satisfy the stationary condition 2.3.

For our procedure to succeed, we must show this primal-dual pair $(\widehat{\theta}, \widehat{Z})$ is optimal for 2.3, in other words

$$\widehat{\theta}_{ij} \neq 0, \qquad \text{for} \qquad (i,j) \in E; \tag{A.95}$$

$$\|\widehat{Z}_{ij}\| < 1, \qquad \text{for} \qquad (i,j) \notin E. \tag{A.96}$$

In the sequel we show these two conditions hold with probability approaching one.

We begin by proving a bound on the remainder.

Lemma A.3.1. The remainder (A.93) satisfies

$$\bar{R} = \left(\|\Delta_E\|_\infty^2 + \|\theta_T^*\|_\infty^2 \right) \bar{K}_{\theta^\dagger}. \tag{A.97}$$

Proof. By Cauchy-Schwarz,

$$\langle \bar{\phi}_E - \mathbb{E}_{\theta^\dagger}[\bar{\phi}_E], \Delta_E \rangle \leq \|\Delta_E\|_\infty \|\bar{\phi}_E - \mathbb{E}_{\theta^\dagger}[\bar{\phi}_E]\|_1, \tag{A.98}$$

and similarly for $\langle \phi_T, \theta_T^* \rangle$,

$$\langle \phi_T - \mathbb{E}_{\theta^\dagger}[\phi_T], \theta_T^* \rangle \leq \|\theta_T^*\|_\infty \|\phi_T - \mathbb{E}_{\theta^\dagger}[\phi_T]\|_1. \tag{A.99}$$

So

$$\bar{R} \leq \max\{\|\Delta\|_\infty, \|\theta_T^*\|_\infty\}^2 \mathbb{E}[(\|\bar{\phi}_E\|_1 + \|\phi_T\|_1)^2 \cdot (\bar{\phi} - \mathbb{E}[\bar{\phi}])]$$

$$= \max\{\|\Delta\|_\infty, \|\theta_T^*\|_\infty\}^2 \bar{K}$$

$$\leq (\|\Delta_E\|_\infty^2 + \|\theta_T^*\|_\infty^2) \bar{K}.$$

Applying the L_∞ norm gives

$$\|\bar{R}\|_\infty \leq (\|\Delta_E\|_\infty^2 + \|\theta_T^*\|_\infty^2)\kappa_R \max\{b(m_2)^2, b(m_1)\}. \tag{A.100}$$

$\square$

Condition (A.95)

Denote $b := \max\{b(m_2)^2, b(m_1)\}$ and $m := \max\{m_2, \sqrt{m_1}\}$.

Lemma A.3.2. Let

$$\tilde{r} := 2\kappa_\Gamma(\|W_E\|_\infty + \lambda_n/m + (\kappa_T + 1)\|\theta_T^*\|_\infty), \tag{A.101}$$

and suppose $\tilde{r} \leq \frac{1}{2b\kappa_R\kappa_\Gamma}$ and $\|\theta_T^*\|_\infty \kappa_R b \leq 1$. Then

$$\|\widehat{\theta}_E - \bar{\theta}_E^*\|_\infty \leq \tilde{r}. \tag{A.102}$$

Proof. The stationary condition for $\widehat{\theta}_E$ is given by

$$\widehat{\mu}_E - \mu(\widehat{\theta})_E + \lambda_n \widehat{Z}_E = 0. \tag{A.103}$$

Denote $W := \widehat{\mu} - \bar{\mu}^*$. Then we may re-write

$$W_E + \bar{\mu}_E^* - \mu(\widehat{\theta})_E + \lambda_n \widehat{Z}_E \tag{A.104}$$

$$= W_E - \Gamma_{EE}\widehat{\Delta}_E - \Gamma_{ET}\theta_T^* - R_E + \lambda_n \widehat{Z}_E, \tag{A.105}$$

where $\widehat{\Delta}_E = \widehat{\theta}_E - \bar{\theta}_E^*$. Re-arranging and applying the L_2 norm, we get for $(i,j) \in E$,

$$\|\widehat{\Delta}_{ij}\|_2 = \|\Gamma_{ij,E}^{-1}(-W_E + \Gamma_{ET}\theta_T^* + R_E - \lambda_n\widehat{Z}_E)\|_2 \tag{A.106}$$

$$= \|\Gamma_{ij,E}^{-1}\|_2(\|W_E\|_2 + m\sqrt{d+E}\|\Gamma_{E,T}\|_\infty\|\theta_T^*\|_\infty + \|R_E\|_2 + \lambda_n\|\widehat{Z}_E\|_2) \tag{A.107}$$

$$\leq \frac{\kappa_\Gamma}{m\sqrt{d+E}}(\|W_E\|_2 + m\sqrt{d+E}\kappa_T\|\theta_T^*\|_\infty + \|R_E\|_2 + \sqrt{d+E}\lambda_n). \tag{A.108}$$

$$= \kappa_\Gamma\left(\|W_E\|_\infty + \kappa_T\|\theta_T^*\|_\infty + \|R_E\|_\infty + \lambda_n/m\right). \tag{A.109}$$

Consider the mapping

$$F(\Delta_E) := -\Gamma_{EE}^{-1}(W_E - \Gamma_{EE}\Delta_E - \Gamma_{EV}\theta_T^* - R_E(\Delta_E) + \lambda_n\widehat{Z}_E) + \Delta_E \tag{A.110}$$

$$= -\Gamma_{EE}^{-1}(W_E - \Gamma_{ET}\theta_T^* - R_E + \lambda_n\widehat{Z}_E). \tag{A.111}$$

Due to the uniqueness of the solution to the stationary conditions, $F(\Delta_E)$ has a unique fixed point $F(\Delta_E) = \Delta_E$ at $\widehat{\Delta} = \widehat{\theta} - (\bar{\theta}^*)$. If we can show that $\|F(\Delta_E)\|_\infty \leq \tilde{r}$ for every $\|\Delta_E\|_\infty \leq \tilde{r}$, since F is continuous and $\{\Delta_E : \|\Delta_E\|_\infty \leq \tilde{r}\}$ is a convex and compact set, applying Brouwer's fixed point theorem [Ortega and Rheinboldt, 2000] implies that the unique fixed point of F satisfies $\|\widehat{\Delta}_E\|_\infty \leq \tilde{r}$. The L_2 norm of the map follows, for $(i,j) \in E$,

$$\|F(\Delta_E)_{ij}\|_2 \leq \|\Gamma_{ij,E}^{-1}\|_2(\|W_E\|_2 + m\|\Gamma_{E,T}\|_\infty\|\theta_T^*\|_\infty + \|R_E\|_2 + \lambda_n\|\widehat{Z}_E\|_2) \tag{A.112}$$

$$\leq \kappa_\Gamma\left(\|W_E\|_\infty + \kappa_T\|\theta_T^*\|_\infty + \|R_E\|_\infty + \lambda_n/m\right). \tag{A.113}$$

Let $\tilde{r} := 2\kappa_\Gamma(\|W_E\|_\infty + (\kappa_T+1)\|\theta_T^*\|_\infty + \lambda_n/m)$ and consider $\|\Delta_{ij}\|_\infty \leq \tilde{r}$ for $(i,j) \in E$. Suppose that also $\tilde{r} \leq \frac{1}{2b\kappa_R\kappa_\Gamma}$.

From lemma A.3.1 the remainder is bounded by

$$\|\bar{R}_E\|_\infty \le b\tilde{\kappa}_R(\|\Delta_E\|_\infty^2 + \|\theta_T^*\|_\infty^2). \tag{A.114}$$

If $\|\Delta_E\|_\infty \le \frac{1}{2b\tilde{\kappa}_R\kappa_\Gamma}$, and $\|\theta_T^*\|_2 b\kappa_R \le 1$, this is bounded by

$$\frac{1}{2\kappa_\Gamma}\|\Delta_E\|_\infty + \|\theta_T^*\|_\infty, \tag{A.115}$$

Thus, since $\|F\|_\infty \le \max_{ij}\|F_{ij}\|_2$,

$$\|F(\Delta_E)\|_\infty \le \kappa_\Gamma\left(\|W_E\|_\infty + (\kappa_\Gamma + 1)\|\theta_T^*\|_\infty + \lambda_n/m + \frac{\|\Delta_E\|_\infty}{2\kappa_\Gamma}\right) \tag{A.116}$$

$$\le \frac{\tilde{r}}{2} + \frac{\tilde{r}}{2} = \tilde{r}. \tag{A.117}$$

It follows that the fixed point $\widehat{\Delta}_E = \widehat{\theta}_E - \bar{\theta}_E^*$ satisfies

$$\|\widehat{\Delta}_E\|_\infty \le 2\kappa_\Gamma(\|W_E\|_\infty + \lambda_n/m + (\kappa_T + 1)\|\theta_T^*\|_\infty). \tag{A.118}$$

$\square$

Condition (A.96)

Lemma A.3.3. Suppose that

$$\max\left\{8\kappa_\Gamma m/\tau\|W\|_\infty, m(\|W\|_\infty + (1 + \kappa_T)\|\theta_T^*\|_\infty)\right\} \le \lambda_n \le \frac{\tau m}{128b\kappa_R\kappa_\Gamma^2}, \tag{A.119}$$

$$\|\theta_T^*\|_\infty \le \min\left\{\frac{\tau\lambda_n}{4(1 + \kappa_T)m}, \frac{m}{b\kappa_R}\right\}, \tag{A.120}$$

Then $\max_{(i,j)\in E^c}\|\widehat{Z}_{ij}\|_2 < 1$.

Proof. Recall that $\widehat{\theta}_{E^c} = \bar{\theta}^*_{E^c} = 0$. For $(i,j) \in E^c$, the stationary condition gives

$$W_{E^c} - \Gamma_{E^c E}((\theta^*_m)_E - \widehat{\theta}_E) - \Gamma_{E^c V}\theta^*_T - R_{E^c} + \lambda_n \tilde{Z}_{E^c} = 0 \qquad (\text{A.121})$$

Now, re-arranging (A.105),

$$\bar{\theta}^*_E - \widehat{\theta}_E = \Gamma^{-1}_{EE}(W_E + \lambda_n \widehat{Z}_E + R_E + \Gamma_{ET}\theta^*_T). \qquad (\text{A.122})$$

It follows that

$$\widehat{Z}_{E^c} = \frac{1}{\lambda_n}\left\{-W_{E^c} - \Gamma_{E^c E}\Gamma^{-1}_{EE}(W_E + \Gamma_{E^c V}\theta^*_T + R_E + \lambda_n\widehat{Z}_E) - \Gamma_{E^c T}\theta^*_T - R_{E^c}\right\}. \qquad (\text{A.123})$$

Now, for $(i,j) \in E^c$,

$$\|\widehat{Z}_{ij}\|_2 \leq \frac{1}{\lambda_n}\left\{\|W_{ij}\|_2 + \|\Gamma_{ij,E}\Gamma^{-1}_{EE}(W_E + \Gamma_V V + R_E + \lambda_n\tilde{Z}_E)\|_2 + \|\tilde{\Gamma}_{ij,E}\theta^*_T\|_2 + \|R_{ij}\|_2\right\} \qquad (\text{A.124})$$

$$\leq \frac{1}{\lambda_n}\left\{\|W_{ij}\|_2 + \|\Gamma_{ij,E}\Gamma^{-1}_{EE}\|_2\left((\|W_E\|_2 + m\sqrt{d+E}\kappa_T\|\theta^*_T\|_\infty + \|R_E\|_2 \right.\right. \qquad (\text{A.125})$$
$$\left.\left. + \lambda_n\|\widehat{Z}_E\|_2\right) + \|R_{ij}\|_2\right\}$$

$$\leq \frac{1}{\lambda_n}\left\{\|W_{ij}\|_2 + \|R_{ij}\|_2 + \frac{(1-\tau)}{\sqrt{d+E}}\left(\|W_E\|_2\right.\right. \qquad (\text{A.126})$$
$$\left.\left. + m\sqrt{d+E}\kappa_T\|\theta^*_T\|_\infty + \|R_E\|_2\right) + (1-\tau)\lambda_n\right\}$$

$$\leq \frac{1}{\lambda_n}\left\{m(2-\tau)(\|W\|_\infty + \|R\|_\infty)\right\} \qquad (\text{A.127})$$
$$+ \frac{1}{\lambda_n}(1-\tau)\kappa_T m\|\theta^*_T\|_\infty + 1 - \tau.$$

From the assumptions of the lemma, and applying Lemma A.3.2, we get

$$\|\Delta_E\|_\infty \le 2\kappa_\Gamma(\|W_E\|_\infty + \lambda_n/m + (\kappa_T + 1)\|\theta_T^*\|_\infty) \qquad (A.128)$$

$$\le 4\kappa_\Gamma \lambda_n/m. \qquad (A.129)$$

From Lemma A.3.1, if $16b\kappa_R\kappa_\Gamma^2\lambda_n \le \frac{\tau m}{8}$ and $b\kappa_R\|\theta_T^*\|_\infty \le m$,

$$\|\bar{R}\|_\infty \le b\kappa_R(\|\Delta_E\|_\infty^2 + \|\theta_T^*\|_\infty^2) \qquad (A.130)$$

$$\le 16b\kappa_R\kappa_\Gamma^2\lambda_n^2/m^2 + m\|\theta_T^*\|_\infty \qquad (A.131)$$

$$\le \frac{\tau\lambda_n}{8m} + m\|\theta_T^*\|_\infty, \qquad (A.132)$$

if $(1 + \kappa_T)m\|\theta_T^*\|_\infty \le \frac{\tau}{4}\lambda_n$, we may bound $\|\widehat{Z}_{ij}\|_2$ by:

$$\|\widehat{Z}_{ij}\|_2 \le (2 - \tau)\frac{\tau}{4} + 1 - \tau + \frac{1 + \kappa_T}{\lambda_n}m\|\theta_T^*\|_\infty \qquad (A.133)$$

$$\le 1 - \tau + \frac{\tau}{2} + \frac{\tau}{4} = 1 - \frac{3\tau}{4} < 1. \qquad (A.134)$$

$\square$

Proof of Theorem 2.5.15. For the assumptions of Lemma A.3.3 to hold, we need

$$\lambda_n = \Omega(m(\|W\|_\infty + \|\theta_T^*\|_\infty)). \qquad (A.135)$$

From Lemma A.2.2, $\|W\|_\infty > t$ with probability no more than

$$2\exp\left\{2\log(nd) - \frac{nt^2}{8\underline{\epsilon}\max\{b(m_2)^4, b(m_1)^2\}}\right\}, \qquad (A.136)$$

116

and the truncation parameters satisfy

$$\|\theta_T^*\|_\infty = O(\max\{m_1^{-r-1/2}, m_2^{-r-1/2}\}). \tag{A.137}$$

Supposing that $m_2 = m_1$, (A.135) is satisfied with the choice

$$\lambda_n \asymp m_2 \left(\sqrt{\frac{m_2^{4\alpha} \log(nd)}{n}} + m_2^{-r-1/2} \right), \tag{A.138}$$

with probability approaching one. Balancing the two terms, if we choose $m_2 \asymp n^{\frac{1}{2r+1+4\alpha}}$, we get

$$\lambda_n \asymp \sqrt{\frac{\log(nd)}{n^{\frac{2r-1}{2r+1+4\alpha}}}} \tag{A.139}$$

Lastly, for condtion A.95 to hold, we need $\max_{(i,j)\in E} \|\widehat{\theta}_{ij} - \bar{\theta}_{ij}^*\|_\infty \leq \frac{\rho^*}{2}$. Using lemma A.3.2, this is satisfied if

$$\frac{\lambda_n}{m_2 \rho^*} \to 0, \tag{A.140}$$

which is satisfied with high probability when

$$\frac{1}{\rho^*} = o \left(\sqrt{\frac{n^{\frac{2r+1}{2r+1+4\alpha}}}{\log(nd)}} \right). \tag{A.141}$$

□

117

APPENDIX B

PROOFS FOR CHAPTER 4

Proof of Proposition 4.2.2. Consider the functional

$$J(q) := \mathbb{E}_p \left[(\|\nabla \log p - \nabla \log q) \otimes x(1-x)\|_2^2 \right].$$

(B.1)

If $J(q) = 0$, then it must be $\nabla \log p = \nabla \log q$ a.e., because their integrated squared distance is zero with respect to a weight function which is nonzero a.e. This implies $\log q = \log p + c$ a.e. for some constant c, but $c = 0$ because p and q must both integrate to one. Furthermore, J is non-negative so it is minimized when $q = p$. If p and q belong to an exponential family with respective natural parameters θ and θ', $\theta = \theta'$ when the family is minimal.

Now,

$$J(q) = \mathbb{E}_p \left[\|\nabla \log q \otimes x(1-x)\|_2^2 \right]$$

(B.2)

$$+ 2\mathbb{E}_p \left[\sum_{i \in V} (\nabla_i \log q \cdot \nabla_i \log p) \otimes x(1-x)\|_2^2 \right] + constant,$$

the constant not depending on q. We have, by integration by parts,

$$\mathbb{E}_p\left[(\nabla_i \log q \nabla_i \log p)x_i(1-x_i)\right] = \int p_i(\nabla_i \log q \nabla_i \log p)x_i(1-x_i) \tag{B.3}$$

$$= \int p_i \frac{\nabla_i p_i}{p_i}(\nabla_i \log q)x_i(1-x_i) \tag{B.4}$$

$$= p_i(x_i)(\nabla_i \log q x_i(1-x_i))\Big]_{x_i=1} \tag{B.5}$$

$$- p_i(x_i)(\nabla_i \log q x_i(1-x_i))\Big]_{x_i=0}$$

$$- \int p_i \nabla_i(\nabla_i \log q x_i(1-x_i))$$

$$= - \int p_i \nabla_i(\nabla_i \log q x_i(1-x_i)), \tag{B.6}$$

where in the last line we applied the boundary assumption. Thus, we see that $J(q)$ is equal to $\mathbb{E}_p\left[h(X,q)\right]$ plus some terms which don't depend on q, so from the argument above $\mathbb{E}_p\left[h(X,q)\right]$ is minimized when $p = q$. We conclude that h is a proper scoring rule.

$\square$

B.1 Parameter Estimation

Lemma B.1.1. If $n \geq Cmd$ and $\lambda_n \geq 2\mathcal{R}^*((\widehat{\Gamma} - \Gamma)\theta^* + \widehat{K} - K)$, with probability at least $1 - 2d \exp\left\{-\frac{\bar{\epsilon}^2}{4\underline{\epsilon}^2}md\right\}$, the regularized score matching estimator $\widehat{\theta}$ satisfies

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{7\lambda_n}{\underline{\epsilon}}\sqrt{d + |E|}. \tag{B.7}$$

Proof. Define the function

$$\mathcal{E}(\delta) = \mathcal{L}(\theta^* + \delta) - \mathcal{L}(\theta^*) + \lambda_n(\mathcal{R}(\theta^* + \delta) - \mathcal{R}(\theta^*)) \tag{B.8}$$

$$= \frac{1}{2}(\theta^* + \delta)^\top \widehat{\Gamma}(\theta^* + \delta) + (\theta^* + \delta)^\top \widehat{K} - \frac{1}{2}(\theta^*)^\top \widehat{\Gamma}(\theta^*)^\top - (\theta^*)^\top \widehat{K} \tag{B.9}$$

$$+ \lambda_n(\mathcal{R}(\theta^* + \delta) - \mathcal{R}(\theta^*))$$

$$= \frac{1}{2}\delta^\top \widehat{\Gamma}\delta + \delta^\top(\widehat{\Gamma}\theta^* + \widehat{K}) + \lambda_n(\mathcal{R}(\theta^* + \delta) - \mathcal{R}(\theta^*)) \tag{B.10}$$

$$= \frac{1}{2}\delta^\top \widehat{\Gamma}\delta + \delta^\top(\widehat{\Gamma}\theta^* - \Gamma\theta^* + \widehat{K} - K) + \lambda_n(\mathcal{R}(\theta^* + \delta) - \mathcal{R}(\theta^*)). \tag{B.11}$$

Since $\mathcal{E}(0) = 0$, it must be that $\mathcal{E}(\widehat{\delta}) \leq 0$.

Using the sub-Gaussian assumption, we may apply [Vershynin, 2010] Remark 5.51, which says for any $c \in (0, 1), t \geq 1$, with probability at least $1 - 2\exp\{-t^2 md\}$, if $n \geq C(t/c)^2 md$, then for any $i \in V$, and any vector δ_i

$$\delta_i^\top \widehat{\Gamma}_i \delta_i \geq \delta_i^\top \Gamma_i \delta_i + \bar{\epsilon}c\|\delta_i\|^2, \tag{B.12}$$

setting $c = \frac{\epsilon}{2\bar{\epsilon}}$, and $t = 1/c$ we get that if $n \geq Cd$, with probability at least $1 - 2\exp\left\{\frac{\bar{\epsilon}^2}{4\underline{\epsilon}^2}d\right\}$,

$$\frac{\epsilon}{2}\|\delta_i\|^2 \leq \delta_i^\top \widehat{\Gamma}_i \delta_i. \tag{B.13}$$

Applying the union bound over all $i \in V$, with probability at least $1 - 2d\exp\left\{\frac{\bar{\epsilon}^2}{4\underline{\epsilon}^2}d\right\}$,

$$\frac{\epsilon}{2}\delta^\top \delta \leq \delta^\top \widehat{\Gamma}\delta, \tag{B.14}$$

where $\delta = (\delta_1^\top, \ldots, \delta_d^\top)^\top$.

By (generalized) Cauchy-Schwarz,

$$\left| \langle \delta, (\widehat{\Gamma} - \Gamma)\theta^* + \widehat{K} - K \rangle \right| \leq \mathcal{R}(\delta)\mathcal{R}^*((\widehat{\Gamma} - \Gamma)\theta^* + \widehat{K} - K) \tag{B.15}$$

$$\leq \frac{\lambda_n}{2}(\mathcal{R}(\delta_{\mathcal{P}}) + \mathcal{R}(\delta_{\mathcal{P}^\perp}). \tag{B.16}$$

where δ_A denotes the projection of δ onto the set A. From [Negahban et al., 2012] Lemma 3, because $\mathcal{R}$ is decomposable, it holds that

$$\mathcal{R}(\theta^* + \delta) - \mathcal{R}(\theta^*) \geq \mathcal{R}(\delta_{\tilde{\mathcal{P}}^\perp}) - \mathcal{R}(\delta_{\tilde{\mathcal{P}}}). \tag{B.17}$$

Combining (B.16) and (B.17),

$$\langle \delta, (\widehat{\Gamma} - \Gamma)\theta^* + \widehat{K} - K \rangle + \frac{\lambda_n}{2}\left(\mathcal{R}(\theta^* + \delta) - \mathcal{R}(\theta^*)\right) \tag{B.18}$$

$$\geq -\frac{3\lambda_n}{2}\mathcal{R}(\delta_{\tilde{\mathcal{P}}}) - \frac{\lambda_n}{2}\mathcal{R}(\delta_{\tilde{\mathcal{P}}^\perp}) \geq -\frac{3\lambda_n}{2}\mathcal{R}(\delta_{\tilde{P}}). \tag{B.19}$$

Using the subspace compatibility constant we have that

$$\mathcal{R}(\delta_{\tilde{\mathcal{P}}}) \leq \sqrt{d + |E|}\|\delta_{\tilde{\mathcal{P}}}\| \leq \sqrt{d + |E|}\|\delta\|. \tag{B.20}$$

Thus conditioning on the aformentioned probability,

$$\mathcal{E}(\delta) \geq \frac{\epsilon}{4}\|\delta\|^2 - \frac{3\lambda}{2}\|\delta\|\sqrt{d + E} \tag{B.21}$$

$$= \|\delta\|\left(\frac{\epsilon}{4}\|\delta\| - \frac{3\lambda_n}{2}\sqrt{d + E}\right). \tag{B.22}$$

Now, consider the set

$$\mathcal{C} = \left\{ \delta : \|\delta\| \leq \frac{7\lambda_n}{\epsilon}\sqrt{d+E} \right\}. \tag{B.23}$$

$\mathcal{C}$ is a compact, convex set. Furthermore, for all $\delta \in \partial\mathcal{C}$, from (B.22) we see that $\mathcal{E}(\delta) > 0$. Also observe that $0 \in \text{int}\mathcal{C}$. Since $\mathcal{E}(\widehat{\delta}) \leq 0$, it must follow that $\widehat{\delta} \in \text{int}\mathcal{C}$, in other words

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{7\lambda_n}{\epsilon}\sqrt{d+E}. \tag{B.24}$$

$\square$

Proof of Theorem 4.5.5. Applying a concentration bound to $\widehat{K}^u_{ij} - K^u_{ij}$ in addition to a union bound, we have that $\|\widehat{K} - K\|_{\max} > t$ with probability no more than $\exp\{2\log(md) - c_2 n t^2\}$ for $t \leq \nu$, for constants $c_2, \nu > 0$. Similarly,

$$\|(\widehat{\Gamma} - \Gamma)\theta^*\|_{\max} \leq 2\kappa_{\theta^*,1}\|\widehat{\Gamma} - \Gamma\|_{\max}, \tag{B.25}$$

and $\|\widehat{\Gamma} - \Gamma\|_{\max} > t$ with probability no more than $\exp\{2\log(md) - c_1 n t^2\}$ for $t < \nu_2$ for $c_1, \nu_2 > 0$. Furthermore, observe that for any vector θ,

$$\mathcal{R}^*(\theta) \leq \sqrt{m}\|\theta\|_{\max}, \tag{B.26}$$

thus setting $\lambda_n \asymp \sqrt{\frac{mk_{1,\theta}^2 \log(md)}{n}}$, the conditions in Lemma B.1.1 will be satisfied with probability approaching one. $\square$

$$\mathcal{C} = \left\{ \delta : \|\delta\| \leq \frac{7\lambda_n}{\epsilon}\sqrt{d+E} \right\}. \tag{B.23}$$

$\mathcal{C}$ is a compact, convex set. Furthermore, for all $\delta \in \partial\mathcal{C}$, from (B.22) we see that $\mathcal{E}(\delta) > 0$. Also observe that $0 \in \text{int}\mathcal{C}$. Since $\mathcal{E}(\widehat{\delta}) \leq 0$, it must follow that $\widehat{\delta} \in \text{int}\mathcal{C}$, in other words

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{7\lambda_n}{\epsilon}\sqrt{d+E}. \tag{B.24}$$

$\square$

Proof of Theorem 4.5.5. Applying a concentration bound to $\widehat{K}^u_{ij} - K^u_{ij}$ in addition to a union bound, we have that $\|\widehat{K} - K\|_{\max} > t$ with probability no more than $\exp\{2\log(md) - c_2 n t^2\}$ for $t \leq \nu$, for constants $c_2, \nu > 0$. Similarly,

$$\|(\widehat{\Gamma} - \Gamma)\theta^*\|_{\max} \leq 2\kappa_{\theta^*,1}\|\widehat{\Gamma} - \Gamma\|_{\max}, \tag{B.25}$$

and $\|\widehat{\Gamma} - \Gamma\|_{\max} > t$ with probability no more than $\exp\{2\log(md) - c_1 n t^2\}$ for $t < \nu_2$ for $c_1, \nu_2 > 0$. Furthermore, observe that for any vector θ,

$$\mathcal{R}^*(\theta) \leq \sqrt{m}\|\theta\|_{\max}, \tag{B.26}$$

thus setting $\lambda_n \asymp \sqrt{\frac{mk_{1,\theta}^2 \log(md)}{n}}$, the conditions in Lemma B.1.1 will be satisfied with probability approaching one. $\square$

B.2   Model Selection

Our proof technique is the *primal dual witness* method, used previously in analysis of model selection for graphical models [Jason D. Lee, 2014; Ravikumar et al., 2011]. It proceeds as follows: construct a primal-dual pair $(\widehat{\theta}, \widehat{Z})$ which satisfies $\mathrm{supp}(\widehat{\theta}) = \mathrm{supp}(\theta^*)$, and also satisfies the stationary conditions for 4.30 with high probability. The stationary conditions for 4.30 are

$$\widehat{\Gamma}\widehat{\theta} + \widehat{K} + \lambda_n \widehat{Z} = 0, \tag{B.27}$$

where $\widehat{Z}$ is an element of the subdifferential $\partial \mathcal{R}(\widehat{\theta}_e)$:

$$(\partial \mathcal{R}(\theta_e))_{ij} = \begin{cases} \{\theta_{ij} : \|\theta_{ij}\|_2 \leq 1\}, & \text{if } \theta_{ij} = 0; \\ \frac{\theta_{ij}}{\|\theta_{ij}} \|_2, & \text{if } \theta_{ij} \neq 0. \end{cases} \tag{B.28}$$

From this we may conclude that there exists a solution to 4.30 is sparsistent. In particular, we have the following steps:

1. Set $\widehat{\theta}_{E^c} = 0$;

2. Set $\widehat{Z}_{ij} = \partial \mathcal{R}(\theta_e^*)_{ij} = \frac{\bar{\theta}_{ij}^*}{\|\theta_{ij}^*\|}$ for $(i,j) \in E$;

3. Given these choices for $\widehat{\theta}_{E^c}$ and $\widehat{Z}_E$, choose $\widehat{\theta}_E$ and $\widehat{Z}_{E^c}$ to satisfy the stationary condition 4.30.

For our procedure to succeed, we must show this primal-dual pair $(\widehat{\theta}, \widehat{Z})$ is optimal for 4.30, in other words

$$\widehat{\theta}_{ij} \neq 0, \qquad\qquad \text{for} \qquad (i,j) \in E; \qquad\qquad \text{(B.29)}$$

$$\|\widehat{Z}_{ij}\| < 1, \qquad\qquad \text{for} \qquad (i,j) \notin E. \qquad\qquad \text{(B.30)}$$

In the sequel we show these two conditions hold with probability approaching one.

Lemma B.2.1. Suppose that $\|\widehat{\Gamma}_{EE} - \Gamma_{EE}\|_{\max} \leq \frac{1}{2ms\kappa_{\Gamma}}$. Then there exists a solution to (4.30), $\widehat{\theta}$, satisfying

$$\|\widehat{\theta}_E - \theta_E^*\|_{\infty} \leq 2\kappa_{\Gamma}\left(2\kappa_{1,\theta}\|(\widehat{\Gamma}_{EE} - \Gamma_{EE})\|_{\max} + \|\widehat{K} - K\|_{\infty} + \lambda_n/\sqrt{m}\right). \qquad \text{(B.31)}$$

Proof. The stationary condition for $\widehat{\theta}_E$, observing that $\widehat{\theta}_{E^c} = \theta_{E^c}^* = 0$, is given by

$$\widehat{\Gamma}_{EE}\widehat{\theta}_E + \widehat{K}_E + \lambda_n\widehat{Z}_E = 0. \qquad\qquad \text{(B.32)}$$

Re-arranging and observing that $\Gamma_{EE}\theta_E^* = -K_E$, we have

$$\widehat{\Gamma}_{EE}\widehat{\theta}_E + \widehat{K}_E + \lambda_n\widehat{Z} = \widehat{\Gamma}_{EE}\widehat{\theta}_E - \Gamma_{EE}\theta_E^* + \widehat{K}_E - K_E + \lambda_n\widehat{Z}_E \qquad\qquad \text{(B.33)}$$

$$= (\widehat{\Gamma}_{EE} - \Gamma_{EE})\widehat{\theta} + \Gamma_{EE}(\widehat{\theta}_E - \theta_E^*) + \widehat{K}_E - K_E + \lambda_n\widehat{Z}_E. \qquad \text{(B.34)}$$

Consider the map

$$F(\Delta_E) = -\Gamma_{EE}^{-1}\left((\widehat{\Gamma}_{EE} - \Gamma_{EE})(\Delta_E + \theta_E^*) + \Gamma_{EE}\Delta_E + \widehat{K}_E - K_E + \lambda_n\widehat{Z}_E\right) + \Delta_E.$$

$$\text{(B.35)}$$

F has a fixed point $F(\Delta_E) = \Delta_E$ at $\widehat{\Delta}_E = \widehat{\theta}_E - \theta_E^*$ for any solution $\widehat{\theta}$. Define $\tilde{r} :=$ $2\kappa_\Gamma \left(2\kappa_{1,\theta} \|(\widehat{\Gamma}_{EE} - \Gamma_{EE})\|_{\max} + \|\widehat{K}_E - K_E\|_\infty + \lambda_n/\sqrt{m} \right)$. If we can show $\|F(\Delta)\|_\infty \leq \tilde{r}$ for each $\|\Delta\|_\infty \leq \tilde{r}$, from Brouwer's fixed point theorem [Ortega and Rheinboldt, 2000], it follows that some fixed point satisfies $\|\widehat{\Delta}\|_\infty \leq \tilde{r}$. For $\|\Delta\|_\infty \leq \tilde{r}$,

$$\|F_{ij}\|_2 \leq \frac{\kappa_\Gamma}{\sqrt{m(d+|E|)}} \left(\|(\widehat{\Gamma}_{EE} - \Gamma_{EE})(\Delta_E + \theta_E^*)\|_2 + \|\widehat{K}_E - K_E\|_2 + \lambda_n\|\widehat{Z}_E\|_2 \right) \tag{B.36}$$

$$\leq \kappa_\Gamma \left(\|(\widehat{\Gamma}_{ij,E} - \Gamma_{ij,E})(\Delta_E + \theta_E^*)\|_\infty + \|\widehat{K}_E - K_E\|_\infty + \lambda_n/\sqrt{m} \right). \tag{B.37}$$

Now,

$$\|(\widehat{\Gamma}_{EE} - \Gamma_{EE})\Delta\|_\infty \leq \left\{ \max_{i \in V} \sum_{j \in V, k \leq m} \left| \Delta_{ij}^k \right| \right\} \cdot \|\widehat{\Gamma}_{EE} - \Gamma_{EE}\|_{\max} \tag{B.38}$$

$$\leq ms\|\Delta\|_\infty \|\widehat{\Gamma}_{EE} - \Gamma_{EE}\|_{\max}, \tag{B.39}$$

and similarly,

$$\|(\widehat{\Gamma}_{EE} - \Gamma_{EE})\theta_E^*\|_\infty \leq 2\kappa_{1,\theta} \|\widehat{\Gamma}_{EE} - \Gamma_{EE}\|_\infty. \tag{B.40}$$

Thus, if $\|\widehat{\Gamma}_{EE} - \Gamma_{EE}\|_{\max} \leq \frac{1}{2ms\kappa_\Gamma}$,

$$\|F\|_\infty \leq \max_{ij} \|F_{ij}\|_2 \leq \kappa_\Gamma \left(\|\widehat{\Gamma}_{EE} - \Gamma_{EE}\|_{\max}(2\kappa_{1,\theta} + ms\tilde{r}) + \|\widehat{K} - K\|_\infty + \lambda_n/\sqrt{m} \right) \tag{B.41}$$

$$\leq \frac{\tilde{r}}{2} + \frac{\tilde{r}}{2} \leq \tilde{r}. \tag{B.42}$$

$\square$

Lemma B.2.2. Suppose that $\sqrt{m}\|\widehat{K} - K\|_\infty \leq \frac{\tau\lambda_n}{4}$, $m^{1/2}\|\widehat{\Gamma} - \Gamma\|_{\max}(\kappa_\theta + s\sqrt{m}\lambda_n) \leq \frac{\tau\lambda_n}{4}$, and $\lambda_n/\sqrt{m} \geq 2\kappa_\Gamma(ms\kappa_\theta\|\widehat{\Gamma}_{EE} - \Gamma_{EE}\|_{\max} + \|\widehat{K} - K\|_\infty)$. Then for each $(i,j) \in E^c$,

$$\|\widehat{Z}_{ij}\|_2 < 1. \tag{B.43}$$

Proof. For $(i,j) \in E^c$, the stationary conditions are

$$0 = \widehat{\Gamma}_{E^cE}\widehat{\theta}_E + \widehat{K}_{E^c} + \lambda_n\widehat{Z}_{E^c} \tag{B.44}$$

$$= \Gamma_{E^cE}(\widehat{\theta}_E - \theta_E^*) + (\widehat{\Gamma}_{E^cE} - \Gamma_{E^cE})\widehat{\theta}_E \tag{B.45}$$

$$+ \widehat{K}_{E^c} - K_{E^c} + \lambda_n\widehat{Z}_{E^c},$$

re-arranging and plugging in the stationary conditions for $\widehat{\theta}_E$, we have for $(i,j) \in E^c$,

$$\widehat{Z}_{ij} = \frac{1}{\lambda_n}\bigg\{ -\Gamma_{ij,E}\Gamma_{EE}^{-1}(-(\widehat{\Gamma}_{EE} - \Gamma_{EE})\widehat{\theta}_E - (\widehat{K}_E - K_E) - \lambda_n\widehat{Z}_E) \tag{B.46}$$

$$- (\widehat{\Gamma}_{ij,E} - \Gamma_{ij,E})\widehat{\theta}_E - \widehat{K}_{ij} + K_{ij}\bigg\}.$$

Applying the L_2 norm,

$$\|\widehat{Z}_{ij}\|_2 = \frac{1}{\lambda_n}\bigg\{ \|\Gamma_{ij,E}\Gamma_{EE}^{-1}\|_2(\|(\widehat{\Gamma}_{EE} - \Gamma_{EE})\widehat{\theta}_E\|_2 + \|\widehat{K}_E - K_E\|_2 + \lambda_n\|\widehat{Z}_E\|_2) \tag{B.47}$$

$$+ \|\widehat{\Gamma}_{ij,E} - \Gamma_{ij,E}\widehat{\theta}_E\|_2 + \|\widehat{K}_{ij} - K_{ij}\|_2\bigg\}$$

$$\leq \frac{1}{\lambda_n}\bigg\{ \sqrt{m}(2-\tau)(\|\widehat{K} - K\|_\infty + \|(\widehat{\Gamma} - \Gamma)\widehat{\theta}_E)\|_\infty\bigg\} + 1 - \tau. \tag{B.48}$$

Observe that since $\|\widehat{\theta} - \theta^*\|_{\max} \leq \tilde{r}$,

$$\|(\widehat{\Gamma} - \Gamma)\widehat{\theta}\|_\infty \leq \|\widehat{\Gamma} - \Gamma\|_{\max}(\kappa_{1,\theta} + ms\tilde{r}) \tag{B.49}$$

$$\leq \|\widehat{\Gamma} - \Gamma\|_{\max}\left(\kappa_{1,\theta} + 2s\sqrt{m}\lambda_n\right), \tag{B.50}$$

so $\|\widehat{Z}_{ij}\|_2$ is bounded by

$$\frac{1}{\lambda_n}\left\{(2-\tau)\sqrt{m}\|\widehat{K} - K\|_\infty + \sqrt{m}(2-\tau)\|\widehat{\Gamma} - \Gamma\|(\kappa_{1,\theta} + 2s\sqrt{m}\lambda_n)\right\} + 1 - \tau. \tag{B.51}$$

if $\sqrt{m}\|\widehat{K} - K\|_\infty \leq \frac{\tau\lambda_n}{4}$ and $\sqrt{m}\|\widehat{\Gamma} - \Gamma\|_{\max}(2\kappa_{1,\theta} + 2s\sqrt{m}\lambda_n) \leq \frac{\tau\lambda_n}{4}$, this is bounded by

$$(2-\tau)\left(\frac{\tau}{4} + \frac{\tau}{4}\right) + 1 - \tau \leq 1 - \frac{\tau}{2} < 1. \tag{B.52}$$

$\square$

Proof of Theorem 4.5.9. Using a concentration bound for $\widehat{K}^u_{ij} - K^u_{ij}$ and applying a union bound, we have that when $t \leq \nu_1$ for some ν, $\|\widehat{K} - K\|_{\max} > t$ with probability no more than $\exp\{2\log(md) - c_2nt^2\}$ for a constant c_2. Similarly,

$$\|(\widehat{\Gamma} - \Gamma)\theta^*\|_{\max} \leq 2\kappa_{1,\theta}\|\widehat{\Gamma} - \Gamma\|_{\max}, \tag{B.53}$$

and for $t \leq \nu_2$, $\|\widehat{\Gamma} - \Gamma\|_{\max} > t$ with probability no more than $\exp\{2\log(md) - c_1nt^2\}$. Thus setting $\lambda_n = C\sqrt{\frac{m\kappa_{1,\theta}^2 \log(md)}{n}}$ for sufficiently large C, and if $\sqrt{\frac{m^2s^2\log md}{n}} = o(1)$, the assumptions of lemma B.2.2 will be satisfied with probability approaching one. Further, assumption (B.29) is satisfied when $\|\widehat{\theta} - \theta^*\|_\infty \leq \frac{\rho^*}{2}$. Since $\|\widehat{\theta} - \theta^*\|_\infty = O(\lambda_n/\sqrt{m})$, we require $\frac{\lambda_n}{\rho^*\sqrt{m}} = o(1)$.

$\square$

Lemma B.2.3. Let ϕ_k be the kth orthonormal Legendre polynomial on $[0,1]$. then

$$\left| x(1-x)\frac{\partial \phi_k(x)}{\partial x} \right| = O(k^{3/2}), \tag{B.54}$$

$$\left| x(1-x)\frac{\partial^2 \phi_k(x)}{\partial x^2} \right| = O(k^{5/2}). \tag{B.55}$$

Proof. From Bonnet's recursion formula [Abramowitz and Stegun, 1964],

$$x(1-x)\frac{d\phi_k(x)}{dx} = \frac{k}{2}\left((2x-1)\phi_k(x) - \sqrt{\frac{2k+1}{2k-1}}\phi_{k-1}(x) \right), \tag{B.56}$$

so taking absolute values of each side, and using $|\phi_k| \le \sqrt{2k+1}$,

$$\left| x(1-x)\frac{d\phi_k}{dx} \right| \le k\left(|\phi_k| + |\phi_{k-1}| \right) \tag{B.57}$$

$$= O\left(k^{3/2} \right). \tag{B.58}$$

Now, using Legendre's differential equation [Abramowitz and Stegun, 1964],

$$4x(1-x)\frac{d^2\phi_k}{dx^2} + 2(2x-1)\frac{d\phi_k}{dx} - k(k+1)\phi_k = 0, \tag{B.59}$$

and using the fact that $\frac{d\phi_k}{dx} \le \frac{k(k+1)\sqrt{2k+1}}{2}$, we find that

$$\left| x(1-x)\frac{d^2\phi_k}{dx^2} \right| \le \frac{1}{2}\left| \frac{d\phi_k}{dx} \right| + \frac{1}{4}k(k+1)\left| \phi_k \right| \tag{B.60}$$

$$= O\left(k^{5/2} \right). \tag{B.61}$$

$\square$

Proof of Theorem 4.5.13. The proof technique is essentially the same as Theorem 4.5.9 so we omit some details. The main difference is that here $K^* = -\Gamma^*\theta^* - \Gamma_T\theta_T$, so we must deal with one additional term in the analysis, the bias from truncation. Suppose $m_1 = m_2$.

We choose λ_n so that with high probability,

$$\lambda_n = \Omega\left(m_2\kappa_{1,\theta}\|\widehat{\Gamma}_{EE} - \Gamma_{EE}\|_{\max} + m_2\|\widehat{K} - K\|_{\max} + \kappa_T m_2\|\theta_T\|_{\max}\right), \qquad \text{(B.62)}$$

$$\lambda_n \to 0, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(B.63)}$$

as well as requiring $n = \Omega(m_2^4 s^2 \log md)$.

Now, applying Lemma B.2.3, $\|A(x)\|_{\max} = O\left(m_2^4\right)$, so applying Hoeffding's inequality and a union bound as well as the boundedness assumption 4.5.12, $\|\widehat{\Gamma}_{EE} - \Gamma_{EE}\|_{\max} \leq C\sqrt{\frac{m_2^8 \log(m_2 d)}{n}}$ with probability approaching one, for sufficiently large constant C. Similarly, $\|\widehat{K} - K\|_{\max} \leq C'\sqrt{\frac{m_2^6 \log(m_2 d)}{n}}$ with probability approaching one. Furthermore, $\|\theta_T\|_{\max} = O(m_2^{-r-1/2})$. Thus, supposing $\kappa_{1,\theta} = O(m_2^2)$, we need

$$\lambda_n \asymp O\left(\sqrt{\frac{m_2^{12} \log(m_2 d)}{n}} + m_2^{-r+1/2}\right). \qquad\qquad \text{(B.64)}$$

Balancing the two terms, we choose $m_2 \asymp n^{\frac{1}{2r+13}}$, so $\lambda_n \asymp \sqrt{\frac{\log(nd)}{n^{\frac{2r-1}{2r+13}}}}$. The stated sample complexity ensures that $\lambda_n \to 0$. Furthermore, $\|\widehat{\theta}_E - \theta_E^*\|_{\max} = O\left(\lambda_n/m_2\right)$, so we require $\frac{\lambda_n}{m_2\rho^*} \to 0$.

$\square$

APPENDIX C

SOFTWARE

Software for Chapter 3 is available at https://github.com/geb5101h/trw. Software for Chapter 4 is available at https://github.com/geb5101h/quasr.

REFERENCES

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Corporation.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2011). Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, pages 19–53.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516.

Barron, A. R. and Sheu, C.-H. (1991). Approximation of density functions by sequences of exponential families. *The Annals of Statistics*, pages 1347–1369.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

Beirlant, J., Dudewicz, E. J., Györfi, L., and Van der Meulen, E. C. (1997). Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39.

Bertsekas, D. P. (1999). *Nonlinear programming*. Athena Scientific.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics*, pages 1071–1095.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Brown, L. D. (1986). Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-monograph series*, pages i–279.

Cai, T., Liu, W., and Luo, X. (2011). A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Cayley, A. (1889). A theorem on trees. *Quart. J. Math*, 23(376-378):69.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons.

Crain, B. R. (1974). Estimation of distributions using orthogonal expansions. *The Annals of Statistics*, 2(3):454–463.

Crain, B. R. (1976). Exponential models, maximum likelihood estimation, and the haar condition. *Journal of the American Statistical Association*, 71(355):737–740.

Crain, B. R. (1977). An information theoretic approach to approximating a probability distribution. *SIAM Journal on Applied Mathematics*, 32(2):339–346.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158.

Dawid, A. P. and Lauritzen, S. L. (2005). The geometry of decision theory. In *Proceedings of the Second International Symposium on Information Geometry and its Applications*, pages 22–28.

Dawid, A. P. and Musio, M. (2014). Theory and applications of proper scoring rules. *Metron*, 72(2):169–183.

Dobruschin, P. (1968). The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability & Its Applications*, 13(2):197–224.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Fan, Y. (1994). Testing the goodness of fit of a parametric density function by kernel method. *Econometric Theory*, 10(02):316–356.

Fix, E. and Hodges, J. (1989). Discriminatory analysis. nonparametric discrimination: consistency properties. *International Statistical Review/Revue Internationale de Statistique*, pages 238–247.

Forbes, P. G. and Lauritzen, S. (2014). Linear estimating equations for exponential families with application to gaussian linear concentration models. *Linear Algebra and its Applications*, In Press.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Gilbert, J. C. and Nocedal, J. (1992). Global convergence properties of conjugate gradient methods for optimization. *SIAM Journal on Optimization*, 2(1):21–42.

Gonzalez, J., Low, Y., and Guestrin, C. (2011). Parallel belief propagation in factor graphs. *Scaling Up Machine Learning: Parallel and Distributed Approaches*.

Gu, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm. *Journal of the American Statistical Association*, 88(422):495–504.

Gu, C. (2002). *Smoothing spline ANOVA models.* Springer.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. In *Journal of Machine Learning Research*, pages 695–709.

Hyvärinen, A. (2007). Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258.

Jason D. Lee, Yuekai Sun, J. E. T. (2014). *On model selection consistency of regularized M-estimators.* arXiv:1305.7477.

Jeon, Y. and Lin, Y. (2006). An effective method for high-dimensional log-density anova estimation, with application to nonparametric graphical model building. *Statistica Sinica*, 16(2):353.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, pages 140–155.

Kakade, S., Shamir, O., Sindharan, K., and Tewari, A. (2010). Learning exponential families in high-dimensions: Strong convexity and sparsity. In *International Conference on Artificial Intelligence and Statistics*, pages 381–388.

Kimeldorf, G. and Wahba, G. (1971). Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95.

Kingma, D. P. and LeCun, Y. (2010). Regularized estimation of image statistics by score matching. In *Advances in neural information processing systems*, pages 1126–1134.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques.* MIT press.

Köster, U., Lindgren, J. T., and Hyvärinen, A. (2009). Estimating markov random field potentials for natural images. In *Independent Component Analysis and Signal Separation*, pages 515–522. Springer.

Kramer, M. A., Eden, U. T., Lepage, K. Q., Kolaczyk, E. D., Bianchi, M. T., and Cash, S. S. (2011). Emergence of persistent networks in long-term intracranial eeg recordings. *The Journal of Neuroscience*, 31(44):15757–15767.

Kullback, S. (1967). A lower bound for discrimination information in terms of variation (corresp.). *Information Theory, IEEE Transactions on*, 13(1):126–127.

Landau, H. J. (1987). *Moments in mathematics*, volume 37. American Mathematical Soc.

Lasserre, J. B. (2007). Semidefinite programming for gradient and hessian computation in maximum entropy estimation. In *Decision and Control, 2007 46th IEEE Conference on*, pages 3060–3064. IEEE.

Lasserre, J.-B. (2009). *Moments, positive polynomials and their applications*, volume 1. World Scientific.

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012a). High dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326.

Liu, H., Wasserman, L., and Lafferty, J. D. (2012b). Exponential concentration for mutual information estimation with application to forests. In *Advances in Neural Information Processing Systems*, pages 2537–2545.

Liu, H., Xu, M., Gu, H., Gupta, A., Lafferty, J., and Wasserman, L. (2011). Forest density estimation. *The Journal of Machine Learning Research*, 12:907–951.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Nemirovsky, A. S. and Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization.* Wiley-Interscience.

Nesterov, Y. (2013). Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161.

Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104.

Neyman, J. (1937). Smooth test for goodness of fit. *Scandinavian Actuarial Journal*, 1937(3-4):149–199.

Noorshams, N. and Wainwright, M. J. (2013). Belief propagation for continuous state spaces: Stochastic message-passing with quantitative guarantees. *Journal of Machine Learning Research*, 14:2799–2835.

Ortega, J. M. and Rheinboldt, W. C. (2000). *Iterative solution of nonlinear equations in several variables*, volume 30. Siam.

Parry, M., Dawid, A. P., Lauritzen, S., et al. (2012). Proper local scoring rules. *The Annals of Statistics*, 40(1):561–592.

Póczos, B., Xiong, L., and Schneider, J. (2012). Nonparametric divergence estimation with applications to machine learning on distributions. *arXiv preprint arXiv:1202.3758*.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, 16(1):356–366.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D., et al. (2010). High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319.

Ravikumar, P., Wainwright, M. J., Raskutti, G., Yu, B., et al. (2011). High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

Reiss, R.-D. (1989). *Approximate distributions of order statistics*. Springer.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Sriperumbudur, B., Fukumizu, K., Kumar, R., Gretton, A., and Hyvärinen, A. (2013). Density estimation in infinite dimensional exponential families. *arXiv preprint arXiv:1312.3516*.

Stam, C. (2010). Use of magnetoencephalography (meg) to study functional brain networks in neurodegenerative disorders. *Journal of the neurological sciences*, 289(1):128–134.

Sudderth, E. B., Ihler, A. T., Isard, M., Freeman, W. T., and Willsky, A. S. (2010). Nonparametric belief propagation. *Communications of the ACM*, 53(10):95–103.

Tibshirani, R. J. et al. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vigário, R., Jousmäki, V., Haemaelaeninen, M., Haft, R., and Oja, E. (1998). Independent component analysis for identification of artifacts in magnetoencephalographic recordings. *Advances in neural information processing systems*, pages 229–235.

Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674.

Wainwright, M. J., Jaakkola, T. S., and Willsky, A. S. (2003). Tree-reweighted belief propagation algorithms and approximate ml estimation by pseudo-moment matching. In *Workshop on Artificial Intelligence and Statistics*, volume 21, page 97. Society for Artificial Intelligence and Statistics Np.

Wainwright, M. J., Jaakkola, T. S., and Willsky, A. S. (2005). A new class of upper bounds on the log partition function. *Information Theory, IEEE Transactions on*, 51(7):2313–2335.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305.

Yang, E., Allen, G., Liu, Z., and Ravikumar, P. K. (2012). Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, pages 1358–1366.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, 11:2261–2286.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.